INSTITUT POLYTECHNIQUE DE PARIS

NNT : 20XXIPPAX020

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

Long-time behaviors of dynamics with mean field interactions

Thèse de doctorat de l'Institut polytechnique de Paris
préparée à l'École polytechnique

École doctorale de mathématiques Hadamard (EDMH) n° 574
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Paris, le 12 juin 2025, par

Songbo Wang

Composition du jury :

Sylvie Méléard Professeure, École polytechnique	Présidente
François Delarue Professeur, Université Côte d'Azur	Rapporteur
Daniel Lacker Associate Professor, Columbia University	Rapporteur absent
Pierre Cardaliaguet Professeur, Université Paris Dauphine-PSL	Examinateur
Arnaud Guillin Professeur, Université Clermont Auvergne	Examinateur
Huyên Pham Professeur, École polytechnique	Examinateur
Zhenjie Ren Professeur, Université Évry Paris-Saclay	Directeur de thèse
Nizar Touzi Professor, New York University	Directeur de thèse absent

Long-time behaviors of dynamics with mean field interactions

Comportements en temps long des dynamiques avec des interactions de champ moyen

Thèse de doctorat
de l'Institut polytechnique de Paris
préparée au centre de mathématiques appliquées
de l'École polytechnique

Songbo Wang
汪嵩博

June 12, 2025
12 juin 2025

Abstract

This thesis is devoted to the study of the long-time behaviors of dynamics with mean field interactions and their associated particle systems. For most cases treated in the thesis, the structural condition for the long-time behaviors is the flat convexity of the mean field energy functional, which is different from the displacement convexity studied in the classical works of optimal transport and gradient flow. The thesis is comprised of three parts. In the first part, we study the overdamped and underdamped mean field Langevin dynamics, which are gradient dynamics associated to a mean field free energy functional, and show their time-uniform propagation of chaos properties by exploiting their gradient structures and a uniform logarithmic Sobolev inequality. In the second part, we first develop some technical results on logarithmic Sobolev inequalities and apply them to get the time-uniform propagation of chaos for various McKean–Vlasov diffusions. Specifically, for the 2D viscous vortex model, we develop strong regularity bounds on its mean field limit on the whole space and show its propagation of chaos by the Jabin–Wang method; we also study its size of chaos problem using the entropy approach of Lacker and obtain time-uniform sharp bounds in the high viscosity regime. In the last part of the thesis, we explore alternative mean field dynamics that originate from convex optimization problems. For the entropy-regularized optimization, we study a fictitious self-play dynamics and a self-interacting diffusion and show their long-time convergences to the solution of the optimization problem. We also consider a nonlinear Schrödinger semigroup, which is a gradient flow for the optimization problem regularized by Fisher information, and show its exponential convergence under a uniform spectral gap condition.

Keywords. Long-time behavior, mean field interaction, propagation of chaos, gradient flow, entropy, logarithmic Sobolev inequality.

Résumé

Cette thèse est consacrée à l'étude des comportements en temps long des dynamiques avec des interactions de champ moyen et des systèmes de particules associés. Pour la plupart des cas traités dans la thèse, la condition structurelle pour les comportements en temps long est la convexité plate de la fonctionnelle d'énergie de champ moyen, qui est différente de la convexité de déplacement étudiée dans les travaux classiques de transport optimal et de flot de gradient. La thèse est composée de trois parties. Dans la première partie, nous étudions les dynamiques de Langevin de champ moyen suramortie et sousamortie, qui sont des dynamiques de gradient associées à une fonctionnelle d'énergie libre de champ moyen, et nous montrons qu'elles présentent des propriétés de propagation du chaos uniforme en temps en exploitant leurs structures de gradient et une inégalité de Sobolev logarithmique uniforme. Dans la deuxième partie, nous développons d'abord quelques résultats techniques sur les inégalités de Sobolev logarithmiques et nous les appliquons pour obtenir la propagation du chaos uniforme en temps pour de diverses diffusions de McKean-Vlasov. En particulier, pour le modèle de vortex visqueux en 2D, nous développons des bornes de régularité fortes sur sa limite de champ moyen sur l'espace entier et nous montrons sa propagation du chaos par la méthode de Jabin-Wang ; nous étudions également son problème de taille du chaos en utilisant l'approche entropique de Lacker et nous obtenons des bornes optimales et uniformes en temps dans le régime de haute viscosité. Dans la dernière partie de la thèse, nous explorons d'autres dynamiques de champ moyen qui proviennent de problèmes d'optimisation convexes. Pour l'optimisation régularisée par l'entropie, nous étudions une dynamique d'auto-jeu fictif et une diffusion auto-interagissante et nous montrons leurs convergences en temps long vers la solution du problème d'optimisation. Nous considérons également un semigroupe de Schrödinger non linéaire, qui est un flot de gradient pour le problème d'optimisation régularisé par l'information de Fisher, et nous montrons sa convergence exponentielle sous une condition de trou spectral uniforme.

Mots-clés. — Comportement en temps long, interaction de champ moyen, propagation du chaos, flot de gradient, entropie, inégalité de Sobolev logarithmique.

Table of Contents

Table des matières

Introduction

The objective of this thesis is to study the long-time behavior of various dynamics with *mean field interactions*. Although it seems difficult to give a both precise and general definition of mean field interactions, we can say that in the scope of this thesis, we are interested in *non-linear flows of probability measures* that are at least formal limits of interacting particle systems, where each particle's equation of motion is influenced by other particles in a more or less equal manner and the total strength of the influences is of order 1. To give a concrete example, consider the *McKean–Vlasov dynamics* described by the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = b(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \text{where } m_t = \operatorname{Law}(X_t). \tag{1}$$

Here the solution X_t to the SDE is supposed to exist on the half line $[0,\infty)$ and takes value in $\mathcal{X}$ with $\mathcal{X}$ being the Euclidean space $\mathbb{R}^d$ or the torus $\mathbb{T}^d$ for some integer $d \geqslant 1$; the drift $b\colon \mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}^d$ is regular enough to ensure the well-posedness of the equation; and B is the standard d-dimensional Brownian motion. To go from the probabilistic view point to the analytic one, we write the evolution partial differential equation (PDE) that should be satisfied by m_t:

$$\partial_t m_t = \Delta m_t - \nabla \cdot \big(b(m_t, \cdot)m\big). \tag{2}$$

The non-linearity of the equation above is due to the dependency on the measure in the drift $b(\cdot,\cdot)$. We say that the non-linearity is of mean field type if we can find a mapping

$$\frac{\delta b}{\delta m}\colon \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$$

that approximates the non-linearity in the sense that

$$b\big((1-t)m + t\delta_y, x\big) - b(m,x) = t\frac{\delta b}{\delta m}(m,x,y) - t\int_{\mathcal{X}} \frac{\delta b}{\delta m}(m,x,z)m(\mathrm{d}z) + o(t)$$

under the limit $[0,1] \ni t \to 0$, for all $m \in \mathcal{P}(\mathcal{X})$ and x, $y \in \mathcal{X}$, where δ_y stands for the Dirac mass at y. This notably excludes local interactions where the drift $b(m,x)$ depends on the local density $m(x)$ of the measure, i.e., $b(m,x) = \beta\big(m(x), x\big)$ for some $\beta\colon \mathbb{R} \times \mathcal{X} \to \mathbb{R}^d$. This also excludes unfortunately the famous Boltzmann model in the kinetic theory, where only particles at the same spatial position are allowed to interact.

We now turn to the particle system that corresponds to the SDE (1) or the Fokker–Planck PDE (2). Let N be an integer $\geqslant 1$ and denote the integer interval $[\![1, N]\!] = \{1, \dots, N\}$ by $[N]$. We introduce the shorthand notation $\boldsymbol{x} \coloneqq$

$(x^1, \dots, x^N) \in \mathcal{X}^N$ for the N-tuple of elements in $\mathcal{X}$ and denote the corresponding empirical measure by

$$\mu^N_{\boldsymbol{x}} \coloneqq \frac{1}{N} \sum_{i \in [N]} \delta_{x^i}.$$

The SDE system of the N particles writes

$$\mathrm{d}X^i_t = b(\mu^N_{\boldsymbol{X}_t}, X^i_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B^i_t, \qquad \text{for } i \in [N], \tag{3}$$

where B^i are independent standard Brownian motions in d dimensions and $\mu^N_{\boldsymbol{X}_t}$ is the empirical measure formed by $\boldsymbol{X}_t = (X^1_t, \dots, X^N_t)$. Switching to the analytic side, we can also write the N-particle evolution PDE for $m^N_t \coloneqq \mathrm{Law}(\boldsymbol{X}_t)$:

$$\partial_t m^N_t = \sum_{i \in [N]} \Delta_i m^N_t - \sum_{i \in [N]} \nabla_i \cdot \bigl(b(\mu^N_{\boldsymbol{x}}, x^i) m^N_t\bigr). \tag{4}$$

The above Fokker–Planck equation is defined on $[0,\infty) \times \mathcal{X}^N$ and is most notably a *linear* equation, as the drift

$$b^{N,i}(\boldsymbol{x}) \coloneqq b(\mu^N_{\boldsymbol{x}}, x^i), \qquad \text{for } i \in [N]$$

is completely determined by the particle configuration $\boldsymbol{x} = (x^1, \dots, x^N)$ and no longer depends on an external probability law. In a way, we have removed the non-linearity from the dynamics at the expense of increasing significantly the dimension of the PDE.

As mentioned above, we expect that when N tends to infinity, the non-linear system described by (1) or (2) provides a good approximation of the N-particle dynamics (3) or (4). More precisely, we expect that if the N particles are initialized independently from m_0, i.e.,

$$\mathrm{Law}(\boldsymbol{X}_0) = m^N_0 = m_0^{\otimes N} = \mathrm{Law}(X_0)^{\otimes N},$$

then the limit

$$\frac{1}{N} \sum_{i \in [N]} \delta_{X^i_t} = \mu^N_{\boldsymbol{X}_t} \to m_t \text{ in probability}, \qquad \text{when } N \to \infty \tag{5}$$

holds for all $t > 0$. This is a *law of large numbers* for interacting particle systems. Moreover, if the particles are exchangeable, i.e., the joint law of the particles does not depend on their ordering, then the convergence of empirical measure above is equivalent to the convergence

$$\mathrm{Law}(X^1_t, \dots, X^k_t) \eqqcolon m^{N,k}_t \to m_t^{\otimes k} \text{ weakly}, \quad \text{when } N \to \infty, \quad \text{for all } k \text{ fixed}, \tag{6}$$

or in other words, the subsystem of k particles is *asymptotically independent* when the size of the whole system tends to infinity. See e.g. Lemma 1.1.2 of Le Bris's thesis [145] for a precise statement of this equivalence. For some historical reasons, the fact that the particles are asymptotically independent from each other is called *chaos* in the early literature of kinetic theory, and thus the mean field limit above is called *propagation of chaos*: once we have chaotic initial condition, then the particles are chaotic for a positive time.

It was discovered by Sznitman [216] in the beginning of 1990s that by using the *synchronous coupling* technique, we can prove propagation of chaos for a large class of drift. We explain the main ideas of this method in this paragraph. The first step of the method is to create N independent copies of the mean field SDE (1), or in analytic terms, consider the N-fold tensorization $m_t^{\otimes N}$ of non-linear flow $m_t = \mathrm{Law}(X_t)$. We denote by $\bar{X}^i$, $i \in [N]$, these independent solutions to the mean field SDE, and by $\bar{B}^i$, $i \in [N]$, the independent Brownian noises driving the dynamics. They satisfy therefore the following SDE

$$\mathrm{d}\bar{X}_t^i = b(m_t, \bar{X}_t^i)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\bar{B}_t^i, \qquad \text{for } i \in [N].$$

We then couple the independent dynamics of $\bar{X}_t^i$ above with the interacting dynamics (3) by demanding the Brownian noises to be exactly the same, or in other words synchronized:

$$B^i = \bar{B}^i, \quad \text{for } i \in [N].$$

Then, by subtracting the SDE of the interacting and the independent particles, we get

$$\mathrm{d}(X_t^i - \bar{X}_t^i) = \big(b(\mu_{\mathbf{X}_t}^N, X_t^i) - b(m_t, \bar{X}_t^i)\big)\,\mathrm{d}t,$$

where the noises are completely cancelled. Now assume that the drift coefficient b is regular enough so that we have the following control:

$$|b(\mu_{\mathbf{X}_t}^N, X_t^i) - b(m_t, \bar{X}_t^i)| \lesssim \frac{1}{N}\sum_{j\in[N]} |X_t^j - \bar{X}_t^j| + |X_t^i - \bar{X}_t^i| + R_t, \tag{7}$$

where R_t is a positive random variable such that $\mathbb{E}[R_t] \to 0$ when $N \to \infty$. In the original work of Sznitman, the drift depends on the measure through a kernel function:

$$b(m,x) = \int_{\mathcal{X}} K(x,y)m(\mathrm{d}y)$$

for some $K\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}^d$. Working with an K that is sufficiently regular, Sznitman showed that the error term R_t corresponds to the error between

$$\frac{1}{N}\sum_{j\in[N]} K(\cdot, \bar{X}_t^j) \quad \text{and} \qquad \int_{\mathcal{X}} K(\cdot, y)m_t(\mathrm{d}y).$$

As $\bar{X}_t^j$, $j \in [N]$, are independent variables of law m_t, the error term R_t can be controlled by $O(N^{-1/2})$ thanks to the classical variance argument and this is the sharp order in N by the central limit theorem. This control can also be verified for b that is jointly Lipschitz continuous in measure and space, where the metric for the measure argument is the *Kantorovich distance* or the L^1-*Wasserstein distance*. Notably, a recent breakthrough of Fournier and Guillin [93] allows us to identify the sharp order in N (which is roughly $O(N^{-1/d})$) for the random error term R_t in the Wasserstein-Lipschitz case. Once the control (7) is established, by taking absolute values and summing over $i \in [N]$, we find

$$\mathrm{d}\sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] \lesssim \sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|]\,\mathrm{d}t + N\,\mathbb{E}[R_t]\,\mathrm{d}t.$$

Then according to Grönwall's lemma, we get

$$\frac{1}{N}\sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] \leqslant \frac{e^{Ct}}{N}\sum_{i\in[N]} \mathbb{E}[|X_0^i - \bar{X}_0^i|] + C\int_0^t e^{C(t-s)}\,\mathbb{E}[R_s]\,\mathrm{d}s.$$

In the case where the initial condition is chaotic $m_0^N = m_0^{\otimes N}$, we can take $\bar{X}_0^i = X_0^i$ so that the first term on the right hand side vanishes. Using the fact that $\mathbb{E}[R_\cdot] \to 0$ when $N \to \infty$, we get

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] = 0$$

and this is sufficient to justify the mean field limit $\mu^N_{\boldsymbol{X}_t} \to m_t$ in (5).

The synchronous coupling method presented above, despite being simple and robust, fails to address the *long-time behavior* of the mean field limit without additional conditions. Indeed, in the case where the error term satisfies the uniform bound

$$\mathbb{E}[R_s] \leqslant \frac{C}{\sqrt{N}},$$

we can only get, upon modifying the constant C,

$$\frac{1}{N} \sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] \leqslant \frac{C(e^{Ct}-1)}{\sqrt{N}}.$$

That is to say, we need an exponentially large number of particles to well approximate the non-linear mean field flow in the long time. This phenomenon is generic in evolutionary systems (recall the Cauchy–Lipschitz theory for ODE) and we must impose structural conditions to avoid such exponential growth of error in time.

Gradient flows and convexities

The main structural condition on the dynamics in the thesis is that the drift is a negative Wasserstein gradient corresponding to a convex *mean field optimization* problem. To be precise, let $F\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be a mean field functional. We say that F admits a Wasserstein gradient $D_m F\colon \mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}^d$ if we have

$$\lim_{t\searrow 0} \frac{F\big((e^{tv})_\# m\big) - F(m)}{t} = \int_{\mathcal{X}} D_m F(m,x) \cdot v(x) m(\mathrm{d}x)$$

for all regular enough measure $m \in \mathcal{P}(\mathcal{X})$ and vector field $v\colon \mathcal{X} \to \mathbb{R}^d$. Here e^{tv} denotes the exponential mapping generated by the vector field that corresponds the ODE $\dot{x} = v(x)$ in the following way:

$$e^{tv} x_0 = x_t, \qquad \text{where } x\colon [0,t] \to \mathcal{X} \text{ solves } \dot{x}_s = v(x_s) \text{ for } s \in [0,t];$$

and $(e^{tv})_\# m$ denotes the pushforward measure of m by the mapping e^{tv}. The structural condition that we imposed above can be precisely stated as follows:

$$b(m,x) = -D_m F(m,x) \qquad \text{for some convex } F\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}. \tag{8}$$

Here, the convexity of F is understood in the following flat interpolation sense:

$$\forall m_0, m_1 \in \mathcal{P}(\mathcal{X}),\ \forall t \in [0,1], \qquad F\big((1-t)m_0 + t m_1\big) \leqslant (1-t)F(m_0) + tF(m_1),$$

and this must not be confused with the displacement convexity, where the interpolation between probability measures is constructed by the optimal transport (see

Chapter 1 for more discussions on the difference between the two notions of convexity). To clarify the ideas, we suppose that the mean field functional F satisfies

$$F(m) = \int_X U(x) m(\mathrm{d}x)$$

for some regular enough potential function $U\colon \mathcal{X} \to \mathbb{R}$. Then the Wasserstein gradient of F is nothing but ∇U (which does not depend on the measure variable) and F is always linear (thus convex) in the flat interpolation sense. But F is displacement convex if and only if the underlying potential U is a convex function (see discussions in [4, Chapter 9]). In this case, the SDE (1), (3) become

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

which is the classical *overdamped Langevin dynamics*. Thus, the mean field dynamics of our interest

$$\mathrm{d}X_t = -D_m F(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \tag{9}$$

is called *overdamped mean field Langevin dynamics* and similarly for its corresponding system of particles. Passing to the analytic side, we can write the associated PDE

$$\partial_t m_t = \Delta m_t + \nabla \cdot \big(D_m F(m_t, \cdot) m_t\big). \tag{10}$$

We mention that the Wasserstein gradient of a mean field functional is also related to its linear functional derivative, whose precise definition will be given in the following chapters. The linear derivative is denoted by $\delta F/\delta m$ and is a mapping from $\mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}$. Under enough regularity, these two derivatives satisfy the following equality:

$$D_m F(m, x) = \nabla_x \frac{\delta F}{\delta m}(m, x).$$

The reason why the condition (8) would lead to long-time properties for the mean field flow (1) is due to a simple yet powerful observation of Jordan, Kinderlehrer and Otto [126] in the late 1990s: the flow of measures associated to the SDE (9) is a *gradient flow* for the *free energy* functional

$$\mathcal{F}(m) = F(m) + H(m), \qquad \text{where } H(m) = \int_{\mathcal{X}} m(x) \log m(x)\,\mathrm{d}x$$

in the L^2-Wasserstein space. Especially, along the flow $t \mapsto m_t$, the free energy $t \mapsto \mathcal{F}(m_t)$ is decreasing. Since the convexity of F ensures that the free energy $\mathcal{F} = F + H$ has a unique minimizer m_*, we can expect that the mean field flow converges to m_*. In other words, the flow (10) provides a dynamical way of solving the optimization problem *regularized by entropy*:

$$\inf_{m \in \mathcal{P}(\mathcal{X})} \mathcal{F}(m) = \inf_{m \in \mathcal{P}(\mathcal{X})} F(m) + H(m). \tag{11}$$

More precisely, denoting the L^2-Wasserstein metric by W_2 and letting h be a time step > 0, we can define iteratively the following discrete flow of probability measures:

$$\mu^h_{n+1} = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \mathcal{F}(\mu) + \frac{W_2^2(\mu, \mu^h_n)}{2h}, \qquad \text{with initial condition } \mu^h_0 = m_0.$$

This discrete scheme is called the *JKO scheme*. And we have the limit

$$\mu^h_{\lfloor t/h \rfloor} \to m_t, \qquad \text{when } h \to 0, \qquad \text{for all } t > 0.$$

The reader can convince oneself on the terminology of "gradient flow" by considering the following analogy in finite dimension. Let n be an integer $\geqslant 1$ and let $V : \mathbb{R}^n \to \mathbb{R}$ be a potential function. The discrete dynamics defined by

$$x^h_{n+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} V(x) + \frac{|x - x^h_n|^2}{2h}, \qquad \text{with initial condition } x^h_0 = x_0$$

is nothing but the implicit Euler scheme

$$x^h_{n+1} = x^h_n - h\nabla V(x^h_{n+1})$$

for the gradient descent

$$\dot{x} = -\nabla V(x),$$

and under regularity assumptions on V, we can show that the discrete dynamics converges to the continuous ODE. Working with displacement convexity, Carrillo, McCann and Villani [39] studied the free energy dissipation and obtained the ergodicity of the non-linear flow (9) in the 2000s. Ambrosio, Gigli and Savaré then translated many of the results obtained under displacement convexity into statements in the abstract formalism of gradient flows in metric spaces, beautifully presented in their monograph [4]. On the other hand, only recently the gradient flow structure and the flat convexity was exploited to obtain long-time behaviors of the overdamped mean field Langevin flow. We mention here the works of K. Hu, Z. Ren, Šiška and Szpruch [117], Nitanda, D. Wu and Suzuki [178], and Chizat [56].

The motivations behind our studies of the flat convexity for the mean field Langevin or the mean field optimization problem are two-fold: theoretical and practical. From the viewpoint of the theory, it is quite natural to try to go beyond the classical literature that relies on the displacement convexity and investigate the alternative convexities that lead to long-time behaviors. Flat convexity is one natural candidate. In fact, interestingly, for mean field game (MFG) systems, which are essentially a pair of a Fokker–Planck and a Hamilton–Jacobi–Bellman equation coupled with each other, the classical condition ensuring the well-posedness on arbitrarily long intervals is the Lasry–Lions monotonicity [143], or the flat convexity in the case of potential games. Somewhat later, Gangbo and Mészáros [94] showed that the displacement convexity is sufficient for the global well-posedness of the MFG. For the practical part, recently there is a growing interest in modelling the training dynamics of neural networks as a gradient flow in the space of probability measures, and in the case of shallow networks, the loss landscape is convex in the flat sense [163, 57, 211, 203]. The reader may refer to the application sections of Chapters 1 and 2 for a detailed introduction to shallow neural networks and their mean field formulation.

Main contributions

One major contribution of this thesis is to study not only the mean field flow (9) in the long time, but also its associated particle system, under the flat convexity of the energy functional F. From the numerical point of view, this is the natural question to raise after long-time behaviors of the mean field limit are established. Indeed, for the neural network example mentioned above, the mean field flow m_t corresponds to the training dynamics of an infinite-neuron network and is not accessible to real computers. The true training dynamics always involve a finite number of particles and are merely approximations to the mean field limit. Recalling that m_t^N is the joint law of the N particles, we wish to show that there exists a time-uniform bound on the approximation error

$$d(m_t^N, m_t^{\otimes N})$$

for some appropriate metric d on the space of probability measures. This property, called *time-uniform propagation of chaos*, is the main goal of the first two chapters, which form Part I of the thesis. In Chapter 1, we develop the ideas explained above and show that the approximation error is uniformly bounded in time. In Chapter 2, we study the kinetic variant of mean field Langevin dynamics and obtain similar results.

Another key ingredient of Part I is the logarithmic Sobolev inequality (log-Sobolev inequality, LSI), and we showcase its importance by summarizing the argument of [178, 56] in the following. As already mentioned above, the method of Part I is based on the gradient structure and the related free energy dissipation of the non-linear Fokker–Planck equation (10). By taking the time-derivative of the free energy functional, we get, at least formally,

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\int_{\mathcal{X}} |\nabla \log m_t + D_m F(m_t, \cdot)|^2 \, \mathrm{d}m_t.$$

Define $\hat{m}$ to be the unique probability measure with density

$$\hat{m}(x) \propto \exp\biggl(-\frac{\delta F}{\delta m}(m, x)\biggr),$$

where $\delta F/\delta m$ is the linear functional derivative of F. Then using the relation between the linear derivative and the Wasserstein gradient, we find

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\int_{\mathcal{X}} \left|\nabla \log \frac{m_t}{\hat{m}_t}\right|^2 \mathrm{d}m_t =: -I(m_t|\hat{m}_t),$$

where the functional $I(\cdot|\cdot)$ is called *relative Fisher information*. The log-Sobolev inequality for the measure $\hat{m}_t$ then allows us to lower bound the Fisher information $I(m_t|\hat{m}_t)$ by the relative entropy

$$H(m_t|\hat{m}_t) \coloneqq \int_{\mathcal{X}} \log \frac{m_t(x)}{\hat{m}_t(x)} m_t(\mathrm{d}x),$$

up to a multiplicative constant. Thanks to the flat convexity, the relative entropy $H(m_t|\hat{m}_t)$ can again be lowered bounded by the relative free energy $\mathcal{F}(m_t) - \inf \mathcal{F}$. Thus by combining the LSI and the convexity, we obtain the exponential contractivity of the free energy.

Being a powerful tool to get exponential contractivity, log-Sobolev inequalities are unfortunately difficult to establish, especially when we do not have direct access to the density of the probability measure concerned. And this is the main objective of the first two chapters of Part II. In Chapters 3 and 4, we provide two class of criteria for the log-Sobolev inequality based on two completely different methods. We then give a few applications of the LSI to long-time behaviors of particle systems in the second half of Chapter 4. Notably, we extend the time-uniform local propagation of chaos of Lacker and Le Flem [142] to the case of non-convex interaction potential. By working with one of the methods more deeply, we manage to prove an L^∞ bound on the Hessian of the log-density of the mean field flow (2) with a Biot–Savart kernel (which contains a singularity) defined on the *whole space*. This technical result allows us to show for the first time the time-uniform propagation of chaos for the 2D vortex model on the whole space. As a continuation of Chapter 4, we still work on the 2D vortex model in Chapter 5 and show its time-uniform sharp local propagation of chaos in the high temperature regime by extending the method of Lacker [140] to singular interactions. The main novelty of our method is that we use a combinatorial technique to solve the hierarchy of entropies which involve additional terms introduced by the singular interaction.

Finally, we move to the last and perhaps the most exotic part of the thesis, where we study long-time behaviors of non-linear dynamics outside the McKean–Vlasov framework that has been discussed till now. Nevertheless, in Chapters 6 and 7, we still focus on the entropy-regularized mean field optimization problem (11). Note that the first-order condition of the problem is equivalent to the fixed-point problem

$$m = \hat{m},$$

where, as we recall, $\hat{m}(x) \propto \exp\bigl(-\frac{\delta F}{\delta m}(m,x)\bigr)$. By interpreting $\hat{m}$ as the best-response strategy to m, the fixed-point formulation can be understood as a Nash equilibrium condition for a self-game where a person plays against himself. Motivated by the fictitious play strategy from the classical game theory, we study following dynamics

$$\partial_t m_t = \alpha(\hat{m}_t - m_t), \qquad \text{for some constant } \alpha > 0$$

called *entropic fictitious play*, and show its convergence to equilibrium in Chapter 6. One major drawback of the entropic fictitious play is that given a player's state m, it is possibly expensive to compute the best response $\hat{m}$ as this usually involves a Monte Carlo run. To overcome this issue, in Chapter 7, we propose a self-interacting diffusion dynamics which can be thought as an intermediate regime between the entropic fictitious play and a linear diffusion process. We will explain this point in more detail below. Convergence to equilibrium for the self-interacting dynamics is also established in the chapter. In the last Chapter 8, we consider instead the mean field optimization problem *regularized by Fisher information*:

$$\inf_{m\in\mathcal{P}(\mathcal{X})} \mathfrak{F}(m) \coloneqq \inf_{m\in\mathcal{P}(\mathcal{X})} F(m) + I(m) \coloneqq \inf_{m\in\mathcal{P}(\mathcal{X})} F(m) + \int_{\mathcal{X}} \frac{|\nabla m(x)|^2}{m(x)}\,\mathrm{d}x, \tag{12}$$

and the associated gradient descent, with the relative entropy measuring the distance between probability measures. In other words, we propose to study the continuous limit of the following JKO scheme

$$\nu^h_{n+1} \coloneqq \operatorname*{argmin}_{\nu\in\mathcal{P}(\mathcal{X})} \mathfrak{F}(\nu) + \frac{H(\nu|\nu^h_n)}{h}, \qquad \text{with initial condition } \nu^h_n = m_0. \tag{13}$$

The resulting dynamics is a non-linear version of the Schrödinger semigroup and is thus called *mean field Schrödinger* dynamics. Its exponential convergence is obtained via a uniform spectral gap, i.e., a uniform Poincaré inequality.

*
* *

In the rest of the introduction, we give detailed technical previews to the eight chapters of the thesis. We discuss also some future prospects in the end.

Preview of Chapters 1 and 2

Uniform propagation of chaos for Langevin

The main results of the two chapters are the time-uniform propagation of chaos for the overdamped and underdamped mean field Langevin dynamics. As the overdamped dynamics has already been defined above in (9), we define here only the underdamped, or the kinetic dynamics

$$\begin{aligned} \mathrm{d}X_t &= V_t\,\mathrm{d}t, \\ \mathrm{d}V_t &= -V_t\,\mathrm{d}t - D_m F(m_t^X, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \text{where } m_t^X = \mathrm{Law}(X_t). \end{aligned} \tag{14}$$

The second-order structure of the dynamics models a Newtonian particle subject to random forces, making it more suitable to describe physical phenomenons. Moreover, the kinetic Langevin dynamics exhibits an analog of Nesterov acceleration for the gradient Markov chain Monte Carlo, that is, the overdamped Langevin. See the work of Y.-A. Ma et al. [157]. The associated N-particle system is defined by duplicating the SDE N times and replacing the dependency on m_t^X by the empirical measure

$$\mu^N_{\boldsymbol{X}_t} = \frac{1}{N}\sum_{i\in[N]} \delta_{X_t^i}.$$

We still denote by m_t^N the joint law of the N particles, but notice that now this law is also joint in space and in speed:

$$m_t^N \coloneqq \mathrm{Law}\big((X_t^1, V_t^1), \ldots, (X_t^N, V_t^N)\big).$$

Our approach to the long-time behaviors of the overdamped and underdamped mean field Langevin is based on the entropic (hypo-)coercivity of the dynamics, which we explain in detail in the following.

Let us focus at the moment on the overdamped case, and denote the unique invariant measure of (2) by m_*. This measure is also the unique minimizer to the mean field free energy functional:

$$m_* = \operatorname*{argmin}_{m\in\mathcal{P}(\mathcal{X})} \mathcal{F}(m) = \operatorname*{argmin}_{m\in\mathcal{P}(\mathcal{X})} F(m) + H(m).$$

Introduce the relative free energy functional

$$\begin{aligned} \mathcal{F}^N(m_t^N|m_*) &\coloneqq \mathcal{F}^N(m_t^N) - N\mathcal{F}(m_*) \\ &\coloneqq N\int_{\mathcal{X}^N} F(\mu^N_{\boldsymbol{x}})m_t^N(\mathrm{d}\boldsymbol{x}) + H(m_t^N) - NF(m_*) - NH(m_*), \end{aligned}$$

and we will consider its evolution in time. Note that in the expression above, we have used the same symbol $H(\cdot)$ for the entropy functional defined for probability measures on $\mathcal{X}^N$ and $\mathcal{X}$. As noted above, the N-particle dynamics is in fact linear, and we have

$$\mathcal{F}^N(m_t^N) = N\int_{\mathcal{X}^N} F(\mu^N_{\boldsymbol{x}})m_t^N(\mathrm{d}\boldsymbol{x}) + H(m_t^N) = H(m_t^N|m_*^N) + \text{constant},$$

where m_*^N is the N-particle invariant measure with density

$$m_*^N(\boldsymbol{x}) \propto \exp\bigl(-NF(\mu_{\boldsymbol{x}}^N)\bigr).$$

So by classical computations, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}^N(m_t^N|m_*) = -\sum_{i\in[N]}\int_{\mathcal{X}^N}\left|\nabla_i \log\frac{m_t^N(\boldsymbol{x})}{m_*^N(\boldsymbol{x})}\right|^2 m_t^N(\mathrm{d}\boldsymbol{x}) =: -I(m_t^N|m_*^N),$$

where $I(\cdot|\cdot)$ stands for the relative Fisher information. The usual approach is then to find conditions for a uniform-in-N LSI for m_*^N so we can conclude as follows:

$$I(m_t^N|m_*^N) \gtrsim H(m_t^N|m_*^N) \geqslant \mathcal{F}^N(m_t^N|m_*) - O(1).$$

This is indeed the method of Malrieu [159] and Carrillo, McCann and Villani [39], and also the more recent works of Guillin, W. Liu, L. Wu, C. Zhang [100, 99]. However, the conditions for the LSI therein seem to be more related to the displacement convexity and do not seem compatible with the flat convexity which is our structural condition for long-time behaviors. Our innovation is to see the N-particle joint Fisher information $I(m_t^N|m_*^N)$ as the average value of Fisher informations between conditional measures of *only one particle*. This is possible as we have the decomposition by component:

$$\begin{aligned}\sum_{i\in[N]}\int_{\mathcal{X}^N}&\left|\nabla_i \log\frac{m_t^N(\boldsymbol{x})}{m_*^N(\boldsymbol{x})}\right|^2 m_t^N(\mathrm{d}\boldsymbol{x})\\ &= \sum_{i\in[N]}\int_{\mathcal{X}^{N-1}}\int_{X}|\nabla_i\log m_t^{N,i|-i}(x^i|\boldsymbol{x}^{-i}) + D_mF(\mu_{\boldsymbol{x}}^N,x^i)|^2\\ &\qquad\qquad m_t^{N,i|-i}(\mathrm{d}x^i|\boldsymbol{x}^{-i})m_t^{-i}(\mathrm{d}\boldsymbol{x}^{-i}).\end{aligned}\tag{15}$$

Here $-i$ stands for the set of all indices except i, namely $[N]\setminus\{i\}$, and $m_t^{N,-i}$, $m_t^{N,i|-i}$ are respectively marginal and conditional measures defined by

$$\begin{aligned}m_t^{N,-i}(\boldsymbol{x}^{-i}) &:= \int_{\mathcal{X}} m_t^N(\boldsymbol{x})\,\mathrm{d}x^i,\\ m_t^{N,i|-i}(x^i|\boldsymbol{x}^{-i}) &:= \frac{m_t^N(\boldsymbol{x})}{m^{N,-i}(\boldsymbol{x}^{-i})}.\end{aligned}$$

By supposing that the probability measures $\hat m$ of the following form

$$\hat m(x) := \frac{\exp\bigl(-\frac{\delta F}{\delta m}(m,x)\bigr)}{\int_{\mathcal{X}}\exp\bigl(-\frac{\delta F}{\delta m}(m,y)\bigr)\,\mathrm{d}y},\qquad \text{where } m\in\mathcal{P}(\mathcal{X})$$

satisfy a uniform LSI, we can (after some manipulations on the measures) apply the *LSI for the one-particle measure* to the conditional Fisher information in the decomposition (15). Here we remark that this componentwise approach is not entirely new, as it was already used to prove the stability of LSI by tensorization (see e.g. [148, Section 5.2]) and the idea of decomposition is in fact the basis of many dimension-free concentration inequalities (see e.g. the discussions on the Efron–Stein inequality in [27, Section 3.1]). The novelty here is that the base measure

m_*^N is not necessarily tensorized and we manage to control the errors that come from the dependency between the particles. Finally, by using the flat convexity of F in the intermediate steps, we obtain an N-particle LSI with an error term:

$$I(m_t^N|m_*^N) \gtrsim \mathcal{F}^N(m_t^N|m_*) - O(1) \tag{16}$$

and this allows us to conclude by Grönwall's lemma

$$\mathcal{F}^N(m_t^N|m_*) \leqslant Ce^{-ct}\mathcal{F}^N(m_t^N|m_*) + O(1) = O(Ne^{-ct} + 1).$$

Note that by replacing m_t^N by the invariant measure m_*^N, we get

$$\mathcal{F}^N(m_*^N|m_*) = O(1)$$

and thus (16) implies

$$I(m_t^N|m_*^N) \gtrsim \mathcal{F}^N(m_t^N|m_*) - \mathcal{F}^N(m_*^N|m_*) - O(1) = H(m_t^N|m_*^N) - O(1),$$

which is a *defective log-Sobolev inequality* for m_*^N with constants uniform in N. For the reason, by an abuse of language, we say that the inequality (16) is a defective (or noised, discretized) version of the mean field non-linear LSI:

$$I(m|\hat{m}) \gtrsim H(m|\hat{m}) \geqslant \mathcal{F}(m) - \mathcal{F}(m_*) \geqslant H(m|m_*). \tag{17}$$

The calculations above will be explained in full detail and rigor in the proof of Theorem 1.12 in Chapter 1. The control on the relative free energy $\mathcal{F}^N(m_t^N|m_*)$ then implies the following, due to the convexity of F and Talagrand's inequality,

$$W_2^2(m_t^N, m_*^{\otimes N}) \lesssim H(m_t^N|m_*^{\otimes N}) = O(Ne^{-ct} + 1).$$

We combine this with the exponential convergence of the mean field flow

$$W_2^2(m_t, m_*) \lesssim H(m_t|m_*) = O(e^{-ct})$$

and the standard propagation of chaos bound that is exponentially growing:

$$\frac{1}{N}W_2^2(m_t^N, m_t^{\otimes N}) = O(e^{ct}N^{-\alpha}), \qquad \text{for some } \alpha > 0.$$

Finally through a triangle argument, we get the desired uniform bound

$$\sup_{t\geqslant 0} \frac{1}{N}W_2^2(m_t^N, m_t^{\otimes N}) = O(N^{-\alpha'}), \qquad \text{for some } \alpha' > 0.$$

We also conduct a rather detailed study of the (reverse) hypercontractivity of the non-linear evolution (10) and obtain a rather strong *lower bound* on the density of m_t in long time. This allows us to mimic the triangle argument above and obtain the uniform bound on the relative entropy:

$$\sup_{t\geqslant 0} \frac{1}{N}H(m_t^N|m_t^{\otimes N}) = O(N^{-\alpha''}), \qquad \text{for some } \alpha'' > 0.$$

In Chapter 2, we turn to the study of the long-time behavior of the hypoelliptic non-linear flow (14) and its associated particle system. Note that in this case, contrary to the overdamped mean field Langevin, the exponential convergence of the mean field flow does not even seem to have been established, and this is also the first contribution of this chapter. By physical intuitions, for a probability measure $m \in \mathcal{P}(\mathcal{X} \times \mathbb{R}^d)$, the natural free energy functional is the sum of the potential and kinetic energies, and the entropy:

$$\mathcal{F}(m) = F(m^X) + \frac{1}{2}\int_{\mathcal{X}\times\mathbb{R}^d} |v|^2 m(\mathrm{d}x\,\mathrm{d}v) + H(m),$$

where m^X is the first marginal of m. The particularity of the hypoellipticity lies in the fact that, when we take the time-derivative of the free energy functional, we do not get a full Fisher information, but only the partial one in the speed directions:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} &= -\int_{\mathcal{X}} |\nabla_v \log m_t(x,v) + v|^2 m_t(\mathrm{d}x\,\mathrm{d}v) \\ &= -\int_{\mathcal{X}} \left|\nabla_v \log \frac{m_t(x,v)}{\hat{m}_t(x,v)}\right|^2 m_t(\mathrm{d}x\,\mathrm{d}v),\end{aligned}$$

where $\hat{m}$ is the probability measure with density

$$\hat{m}(x,v) \propto \exp\biggl(-\frac{\delta F}{\delta m}(m^X, x) - \frac{1}{2}|v|^2\biggr).$$

One solution to deal with this degeneracy common in kinetic models is to introduce a distorted quantity, as demonstrated by Y. Guo [103] and Talay [217], and this is indeed the solution that Villani found for the linear Fokker–Planck dynamics [221, Part I]. In the linear case, the potential energy F is given by a potential function:

$$F(m^X) = \int_{\mathcal{X}} U(x) m^X(\mathrm{d}x), \qquad \text{for some } U\colon \mathcal{X} \to \mathbb{R},$$

and thus $\hat{m}$ is independent of m and is identical to the equilibrium measure, which we denote by m_*. The idea of Villani is to introduce an anisotropic (but always positive-definite) Fisher information

$$\begin{aligned}I_{a,b,c}(m|m_*) = \int_{\mathcal{X}\times\mathbb{R}^d} \biggl(a\left|\nabla_v \log \frac{m(z)}{m_*(z)}\right|^2 &+ 2b\nabla_v \log\frac{m(z)}{m_*(z)} \cdot \nabla_x \log \frac{m(z)}{\hat{m}(z)} \\ &+ c\left|\nabla_x \log\frac{m(z)}{m_*(z)}\right|^2\biggr) m_t(\mathrm{d}z),\end{aligned}$$

where $z = (x,v)$. By choosing the right a, b, c and doing some lengthy computations, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigl(\mathcal{F}(m_t) + I_{a,b,c}(m_t|m_*)\bigr) \lesssim -\int_{\mathcal{X}\times\mathbb{R}^d} \left|\nabla_z \log\frac{m_t(z)}{m_*(z)}\right|^2 m_t(\mathrm{d}z) =: -I(m_t|m_*).$$

The full Fisher information on the right-hand side allows us to conclude by the usual LSI for m_* and the phenomenon is called *hypocoercivity*. In our work, we adapt this construction of Villani to our non-linear setting, by replacing the m_* in the anisotropic Fisher by $\hat{m}_t$. We then need to compute the time-derivative of

$$\mathcal{F}(m_t) + I_{a,b,c}(m_t|\hat{m}_t).$$

Most of the computations follow in line with Villani's original ones since the generator of the evolution at time t annihilates the measure $\hat{m}_t$, but we also need to control the additional term that comes from the variation of $\hat{m}_t$. This fortunately does not pose any problem to the hypocoercivity and we obtain the exponential convergence of the mean field flow (see also the proof of Theorem 2.2).

For the associated N-particle system, since the dynamics is linear, we can directly apply the formalism of Villani, and the only important point is to have a hypocoercivity that is uniform in N. To be precise, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(H(m_t^N|m_*^N) + I_{a,b,c}(m_t^N|m_*^N)\big) \leqslant -\kappa I(m_t^N|m_*^N)$$

for some a, b, c and $\kappa > 0$ that do not depend on N. In the kinetic case, we also have a discretized LSI that is similar to (16), and this allows us to get again

$$\mathcal{F}^N(m_t^N) - N\mathcal{F}(m_*) \leqslant O(Ne^{-ct} + 1).$$

The arguments are explained in detail in the proof of Theorem 2.3.

Finally, we also work on the short-time regularization properties of the hypoelliptic dynamics. By adapting the *coupling by change of measure* method developed by Guillin, P. Ren and F.-Y. Wang in a series of works [102, 227, 193], we obtain dimension-free log-Harnack inequalities for the mean field and the particle system flows, which lead to the regularization from Wasserstein distance to relative entropy. Then we adapt Hérau's functional [109] again to our setting to obtain the regularization from relative entropy to relative Fisher information. We finally combine the long-time exponential convergences, short-time regularizations and standard exponentially growing propagation of chaos to derive a time-uniform propagation of chaos bound, without requiring any regularity of the initial data (see Theorem 2.6).

Preview of Chapters 3 and 4

LSI and applications

The main objective of Chapters 3 and 4 is to obtain LSI criteria for probability measures related to a diffusion process without directly having access to their densities. In Chapter 3, we wish to show an LSI for the *stationary* measure to the following time-homogeneous diffusion

$$\partial_t m_t = \frac{\sigma^2}{2}\Delta m_t - \nabla\cdot(bm_t),$$

and in Chapter 4, we allow the drift b to depend on time and wish to show a *time-uniform* LSI for m_t solving the inhomogeneous

$$\partial_t m_t = \frac{\sigma^2}{2}\Delta m_t - \nabla\cdot(b_t m_t). \tag{18}$$

First note that the first problem is in fact more or less included in the second one, since if we can show a time-uniform LSI for m_t, $t \geqslant 0$, and we know that m_t converges to some stationary measure m_* weakly, then m_* also satisfies an LSI. So we will primarily work within the second parabolic framework in the rest of the preview. Secondly, if the drift in the first problem is a gradient:

$$b(x) = -\nabla U(x),$$

then the diffusion is symmetric and we know that the stationary measure has density proportional to $\exp\bigl(-2U(x)/\sigma^2\bigr)$. In this case, by the classical Bakry–Émery criterion and perturbation results, we already know how to show LSI for a large class of U. Hence, the interest of Chapter 3 lies in the non-symmetric case. As a final note, in the case where b satisfies a strong monotonicity:

$$\bigl(b_t(x) - b_t(y)\bigr) \cdot (x-y) \leqslant -\kappa|x-y|^2, \qquad \text{for some } \kappa > 0,$$

then by repeating the classical argument of Malrieu [159] (see also the beginning of Chapter 4), we can propagate the LSI uniformly in time, and the second problem is solved. So we will focus on cases where such strong monotonicity is absent.

As mentioned above, we provide two class of such criteria based on two different methods. The first method is based on a recent work of Monmarché [168], where he showed that, if b is regular enough and if b is only non-monotone inside a compact set, then there exists σ_0 such that for all $\sigma \geqslant \sigma_0$, the diffusion process related to (18) is contractive in L^2-Wasserstein. In other words, for two solutions μ_t, ν_t to (18), we have

$$W_2(\mu_t, \nu_t) \leqslant Me^{-\lambda t} W_2(\mu_0, \nu_0), \qquad \text{for some } \lambda > 0. \tag{19}$$

Two proofs of this result are given in the cited work of Monmarché and the probabilistic proof is based on the synchronous coupling of the diffusion processes and a modified transport cost that is equivalent to the squared Euclidean distance. The high temperature condition is crucially used to construct such a transport cost. Then, by standard arguments (in fact, an L^2 version of Malrieu's propagation result [159]), the contraction (19) leads to a uniform Poincaré inequality for m_t. We then observe that, once we have some time-uniform control on $x \mapsto x \cdot b_t(x)$ and ∇b_t, we can obtain respectively a Gaussian moment bound and a Harnack inequality (of F.-Y. Wang) that are both uniform in time. These two results lead to a time-uniform hypercontractivity, which is equivalent to a time-uniform defective log-Sobolev inequality. Then by combining the Poincaré and the defective LSI, we get the desired time-uniform LSI.

The second method is based on direct estimates on the density of the solution m_t to (18) and may seem brutal to readers versed in functional inequalities. To simplify we fix $\sigma = \sqrt{2}$ in this paragraph. Suppose that we have a reference measure μ_0 satisfying an LSI which is also stationary to the drift a_0, that is to say,

$$\Delta\mu_0 - \nabla \cdot (a_0\mu_0) = 0.$$

Denote the difference of drifts by $g_t \coloneqq b_t - a_0$ and the log-relative density by $u_t \coloneqq \log m_t/\mu_0$. According to the respective PDE of m_t and μ_0, we find that u_t solves the Hamilton–Jacobi–Bellman (HJB) equation

$$\partial_t u_t = \Delta u_t + |\nabla u_t|^2 + \tilde{b}_t \cdot \nabla u_t + \varphi_t,$$

where the coefficients are defined by

$$\begin{aligned} \tilde{b}_t &\coloneqq 2\nabla \log \mu_0 - b_t, \\ \varphi_t &\coloneqq -\nabla \cdot g_t + g_t \cdot \nabla \log \mu_0. \end{aligned}$$

We say that the drift $\tilde{b}$ is *weakly semi-monotone* if

$$\bigl(\tilde{b}_t(x) - \tilde{b}_t(y)\bigr) \cdot (x-y) \leqslant -\kappa(|x-y|)|x-y|^2,$$

for some $\kappa\colon (0,\infty) \to \mathbb{R}$ such that $\liminf_{r\to\infty} \kappa(r) > 0$ and $r \mapsto r|\kappa(r)|$ is integrable near 0. In a recent work [61], Conforti showed that if $\tilde{b}$ is weakly semi-monotone and φ is Lipschitz continuous, then we have a time-uniform gradient bound on u_t. Thus, according to the log-Lipschitz perturbation result of Aida and Shigekawa [1], we know that $m_t \propto \mu_0 \exp(u_t)$ satisfies a time-uniform LSI. Conforti's method for this gradient estimate is probabilistic: he uses the *coupling by reflection* for controlled diffusion processes and shows the contraction in W_1 distance, which leads to the time-uniform gradient estimate. It seems that the uncontrolled version of coupling by reflection was first developed by Lindvall and Rogers [152] in the 1980s. This coupling was generalized to diffusions on manifolds by Kendall [130] and then used to derive gradient estimates for the heat equation by Cranston [65]. The weak semi-convexity, along with reflection coupling, was exploited by M.-F. Chen and F.-Y. Wang [47] in the 1990s to estimate the spectral gap of the diffusion generator and short-time regularizing effects are derived by Priola and F.-Y. Wang in [187]. Porretta and Priola then showed the regularization effect for the non-linear HJB flow in [186] by using the purely analytical comparison principle between viscosity solutions. The more recent work of Eberle [83] revived this method as it drew a lot of attention from the statistics and machine learning communities. We remark that the above-cited work of Conforti made two contributions that are vital to Chapter 4: first, the *long-time* estimate in the HJB case is obtained; and second, the Hessian estimate (i.e. on $\nabla^2 u_t$) is also proved. We will comment in particular on the second contribution in below.

In the rest of Chapter 4, we discuss a few examples that verify the two criteria presented above and applied the time-uniform log-Sobolev to obtain time-uniform sharp local propagation of chaos for McKean–Vlasov dynamics with non-convex interaction potentials, which is not included in the paper of Lacker and Le Flem [142]. The sharp local propagation of chaos will be discussed in more detail in the preview of the next chapter. However, the most interesting application of our method is perhaps the *2D vortex model on the whole space* presented in the end of the chapter. The 2D vortex model is a probabilistic formulation of the 2D incompressible Navier–Stokes equations and we refer the reader to the expository article [205] for more details. In this model, the mean field flow (1) follows the McKean–Vlasov drift

$$b(m,x) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{(x-y)^\perp}{|x-y|^2} m(\mathrm{d}y),$$

where the symbol $\perp$ stands for the rotation in 2D: $(x_1,x_2)^\perp = (-x_2,x_1)$. In other words, we have that $b(m,x) = (K \star m)(x)$ where K is the Biot–Savart kernel. Recently, Jabin and Z. Wang [124] showed global-in-time propagation of chaos for this model and Guillin, Le Bris and Monmarché [98] improved it into a uniform propagation of chaos bound. However, since the method of Jabin and Z. Wang is based on a *weak-strong uniqueness principle*, it requires rather strong regularity on the mean field flow. To be more precise, one needs to control the L^∞ norm of $\nabla \log m_t$ and $\nabla^2 \log m_t$. This is rather difficult to establish on the whole space as we cannot have a global lower bound on the density m_t. For this reason, the two works cited above only treat the periodic 2D vortex model on the two-dimensional torus. In Chapter 4, we show that by adding an additional quadratic confinement, i.e., by letting

$$b(m,x) = -\kappa x + (K \star m)(x),$$

we can use the HJB method to get L^∞ bounds on

$$\nabla \log \frac{m_t}{m_*},\ \nabla^2 \log \frac{m_t}{m_*}, \qquad \text{where } m_* \propto \exp\biggl(-\frac{\kappa|x|^2}{2}\biggr).$$

Moreover, these bounds converge to zero exponentially fast and this allows us to show the *generation of chaos* property for the 2D vortex model. The proofs of such L^∞ bounds are moderately lengthy. Due to the singularity of the Biot–Savart kernel K, we rely on a parabolic bootstrap procedure to gradually gain regularity on the coefficients $\tilde{b}$, φ, and we need both the long-time contraction of Conforti and the short-time regularization of Porretta, Priola and F.-Y. Wang. To conclude the discussion on 2D vortex, we mention that, after Chapter 4 appeared as preprint [170], Rosenzweig and Serfaty uploaded their preprint [201] where they showed that the 2D vortex models with and without quadratic confinement are equivalent up to a scaling transform. Thus our method can also be applied to the model on the whole space without confinement.

Finally, we mention that in Chapter 3, we also develop a LSI criterion for the stationary measure of a kinetic diffusion. It is based on the HJB method and the most important step is to construct a Wasserstein contraction for the controlled kinetic diffusion processes. The statement and proof are presented in the end of the chapter. The method is based on a mixed coupling, comprising both synchronous and reflective parts, and a distorted (usual for kinetic models) transport cost motivated by the construction of Eberle, Guillin and Zimmer [84]. This result can be considered as a generalization (or even an improvement in certain aspects) to the recent works of Kazeykina, Z. Ren, X. Tan and J. Yang [128] and Schuh [206].

Preview of Chapter 5

Size of chaos for singular dynamics

In Chapter 5, we study a fine property of the mean field large particle system called *local propagation of chaos*. Although we have studied the quantitative propagation of chaos in the previous chapters, the results obtained only concerns distances between the particle and the tensorized mean field system as a whole, for example, the Wasserstein distance $W_2^2(m_t^N, m_t^{\otimes N})$ or the relative entropy $H(m_t^N|m_t^{\otimes N})$. In this chapter, instead of studying these *global* distances, we only observe the first k particles from the N-particle system (3) and compare it with the k-tensorized mean field flow. To justify considering only the first k particles and not other sets of k particles, we need to of course suppose the exchangeability in the N particle system, and this assumption will be in force throughout the chapter. We will also assume that the mean field interaction in the drift takes the following kernel form:

$$b(m,x) = \int_{\mathcal{X}} K(x,y)m(\mathrm{d}y).$$

Recall that the law of the subsystem of k particles at time t is denoted by by $m_t^{N,k}$, or in other words,

$$m_t^{N,k} \coloneqq \mathrm{Law}(X_t^1, \dots, X_t^k),$$

where the dependency of N on the right-hand side is implicit. So the question raised above consists of finding a sharp bound between the two probability measures $m_t^{N,k}$ and $m_t^{\otimes k}$. This is an quantitative version of the chaos condition (6).

For this reason, this question is also called *size of chaos* in the recent literature. See the works of Paul, Pulvirenti and Simonella [182], Duerinckx [78], and Bernou and Duerinckx [20]. One common point of the approaches of these works is that the authors decompose the law of the N particles into a combinatorial sum of connected correlation functions (or cumulants), the decomposition being called the cluster expansion, and they study the evolution of correlation functions along the dynamics. Depending on the specific mean field interaction, correlations between the particles can be generated by collisions or through the drift and the randomness may come from the initialization and also the dynamical noise. Then after estimating the size of the cumulants, they go back to the problem of size of chaos and get

$$\|m_t^{N,k} - m_t^{\otimes k}\| = O\left(\frac{k^2}{N}\right),$$

where $\|\cdot\|$ denotes an appropriate functional norm. Roughly speaking, the k^2 factor arises from counting the number of pairs among the first k particles. This factor cannot be reduced within this approach, unless some cancellation occurs, which may result from orthogonality.

A completely different approach is developed in Lacker's recent work [140], where he considers directly the evolution of the errors between $m_t^{N,k}$ and $m_t^{\otimes k}$, measured in terms of relative entropy

$$H_t^k \coloneqq H(m_t^{N,k}|m_t^{\otimes k}).$$

The dynamics of the measure $m_t^{N,k}$ is described by the BBGKY hierarchy and involves the next-order marginal $m_t^{N,k+1}$, and so the evolution of H_t^k should also involve next-order quantities. Indeed, in the final step of Lacker's proof, the dynamical equation reads

$$\frac{\mathrm{d}H_t^k}{\mathrm{d}t} \leqslant M\frac{k^2}{N^2} + M(H_t^{k+1} - H_t^k),$$

where M is a constant related to the strength of the mean field interaction kernel K. Then solving the system of inequalities above yields the global-in-time bound $H_t^k = O(k^2/N^2)$, and in terms of norm distance,

$$\|m_t^{N,k} - m_t^{\otimes k}\|_{\mathrm{TV}} = O\left(\frac{k}{N}\right)$$

which improves the results above by a factor of k. This bound is sharp as it can be attained by a simple Gaussian example. Later on this method was extended to the time-uniform case in the weakly interacting regime by Lacker and Le Flem [142] and the sharp bound for higher-order chaos is obtained by Hess-Childs and Rowan [111]. We note that the method of Lacker crucially relies on the Brownian noise to control the growth of H_t^k and this is possibly the reason for the gain of factor k compared to the more combinatorial approaches above.

One common limitation of the previous works on the size of chaos is that we require a strong regularity assumption (at least L^∞) on the interaction kernel K and thus excluding the interesting 2D vortex model where the kernel is Biot–Savart:

$$K(x,y) = \frac{(x-y)^\perp}{2\pi|x-y|^2}.$$

The aim of Chapter 5 is precisely to overcome this limitation and show the sharp size of chaos bound for the 2D vortex particle system, that is to say, $H_t^k = O(k^2/N^2)$. By combining the techniques of Jabin–Z. Wang and Lacker, we show that the evolution of the relative entropy verifies

$$\frac{\mathrm{d}H_t^k}{\mathrm{d}t} \leqslant -c_1 I_t^k + c_2 I_t^{k+1} + M\left(H_t^k + \frac{k^2}{N^2}\right) + M(H_t^{k+1} - H_t^k),$$

where I_t^k, I_t^{k+1} are relative Fisher informations. In particular, we have

$$I_t^k \coloneqq \sum_{i\in[k]} \int_{\mathcal{X}^k} \left|\nabla_i \log \frac{m_t^{N,k}(\boldsymbol{x}^{[k]})}{m_t^{\otimes k}(\boldsymbol{x}^{[k]})}\right|^2 m_t^{N,k}(\mathrm{d}\boldsymbol{x}^{[k]}).$$

The main difference compared to Lacker's work is of course the additional *positive Fisher information of the next order* I_t^{k+1}, which comes from the singularity of the kernel K. Solving the system of differential inequalities in the case $c_2 < c_1$ is the main technical innovation of the chapter. We note that the condition $c_2 < c_1$ corresponds to the fact that the $W^{-1,\infty}$ norm of the kernel K is smaller than 1, so our result is valid for weak vortex interactions, or equivalently vortices in the high temperature regime. The main idea of the proof is to consider a weighted mix of entropies of order $\geqslant k$:

$$Z_t^k \coloneqq \sum_{i=k}^{N} a_{k,i} H_t^i, \qquad \text{where } a_{k,i} \geqslant 0 \text{ and } a_{k,k} = 1$$

By choosing the appropriate coefficients $a_{k,i}$, we can cancel all the Fisher informations in the dynamics of Z_t^k and recover the original system of Lacker. So we deduce $Z_t^k = O(k^2/N^2)$ and we can conclude by $H_t^k \leqslant Z_t^k$. Using the ideas from [98] we also improve the global-in-time size of chaos bound into a uniform one. Some consequences are also discussed. For example, by leveraging the injection from L^d into $W^{-1,\infty}$ [28], we can show a global-in-time sharp size of chaos bound for L^d interactions of any strength. We also use an L^2 (instead of entropy) approach for the size of chaos in the vortex interaction case in order to lift the restriction on the interaction strength, but unfortunately only a finite-time result is obtained.

Preview of Chapters 6 and 7

Fictitious play and self-interaction

In Chapters 6 and 7, we study alternative mean field dynamics that approach the minimizer of the entropy regularized mean field optimization problem (11) in long time. The dynamics of interest in Chapter 6 is the entropic fictitious play defined in the following way:

$$\partial_t m_t = \alpha(\hat{m}_t - m_t), \qquad \text{where } \hat{m}_t \propto \exp\left(-\frac{\delta F}{\delta m}(m,\cdot)\right). \tag{20}$$

The definition of the dynamics above is motivated by the fictitious play algorithm, first proposed by Brown [34] in the framework of a two-person game. In a symmetric two-person game with continuous state space, we denote the state of the two players by x, y respectively and the Nash equilibrium condition writes

$$x_* \in \mathrm{BR}(y_*), \qquad y_* \in \mathrm{BR}(x_*),$$

where $\mathrm{BR}(\cdot)$ is the set of best response given the state of the adversary. Brown proposes that both players follow the respective discrete dynamics

$$\begin{aligned} x_{t+1} &= \frac{t}{t+1} x_t + \frac{1}{t+1} a_t, \qquad \text{where } a_t \in \mathrm{BR}(y_t), \\ y_{t+1} &= \frac{t}{t+1} y_t + \frac{1}{t+1} b_t, \qquad \text{where } b_t \in \mathrm{BR}(x_t), \end{aligned}$$

and expects that (x_t, y_t) converges to some Nash equilibrium (x_*, y_*) in long time. To see the intuitions behind our entropic fictitious play dynamics, note that by variational calculus, the first-order condition of the optimization problem (11) reads

$$\frac{\delta F}{\delta m}(m_*, x) + \log m_*(x) = \text{constant}.$$

According to the definition the measure $\hat{m}$, the above condition is equivalent to

$$m_* = \hat{m}_*.$$

It is a Nash equilibrium condition for the one-person (or self) game if the mapping $m \mapsto \hat{m}$ is interpreted as the best-response mapping. And if we replace the $1/t$ scaling in Brown's dynamics by an exponentially scaling, and consider the continuous-time version, the corresponding fictitious play dynamics is exactly (20). We note that Cardaliaguet and Hadikhanloo also used this idea in order to find solution to mean field games [36], which can also be formulated as a fixed-point problem.

The convergence of the fictitious play algorithm is in general not guaranteed, but in the case of potential games, we can usually find Lyapunov functions which decreases along the dynamics. For our entropic fictitious play, we calculate the time-derivative of the free energy functional (which is the functional to optimize) along the dynamics, and find

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\alpha\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big).$$

Since we always suppose that the energy function F is flat-convex, at this point we can already use the entropy sandwich inequality (1.40) in Chapter 1:

$$H(m_t|\hat{m}_t) \geqslant \mathcal{F}(m_t) - \mathcal{F}(m_*)$$

and the exponential convergence follows. This is however not the approach that we take in Chapter 6, partly because we were not familiar with such sandwich inequality when the corresponding paper [49] was written. (This sandwich inequality was already used in [56, 178] to show the exponential convergence of the overdamped mean field Langevin dynamics at that time.) Instead, we take again the time-derivative of $H(m_t|\hat{m}_t)$ and find that at least formally,

$$\begin{aligned} \frac{\mathrm{d}H(m_t|\hat{m}_t)}{\mathrm{d}t} &= -\alpha\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big) \\ &\qquad - \alpha \iint_{\mathcal{X}^2} \frac{\delta^2 F}{\delta m^2}(m_t, x, y)(\hat{m}_t - m_t)^{\otimes 2}(\mathrm{d}x\,\mathrm{d}y). \end{aligned}$$

The last term is negative by the convexity of F, so we get that

$$\frac{\mathrm{d}H(m_t|\hat{m}_t)}{\mathrm{d}t} \leqslant -\alpha H(m_t|\hat{m}_t),$$
$$\frac{\mathrm{d}H(m_t|\hat{m}_t)}{\mathrm{d}t} \leqslant \frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t}.$$

And by some elementary calculus we conclude that $\mathcal{F}(m_t)$ decreases exponentially. The second-order calculation above is interesting by itself as it resembles the Otto–Villani proof [180] of the Bakry–Émery criterion, which we sketch here in a minimalist style. Suppose we are given an overdamped Langevin dynamics, generated by $\Delta - \nabla U \cdot \nabla$ for some $U \colon \mathbb{R}^d \to \mathbb{R}$ satisfying $\nabla^2 U \succeq \rho$ with $\rho > 0$. Denote the invariant measure proportional to $\exp(-U)$ by m_* and let m_t be the associated flow of measure. Denote also for simplicity $H_t \coloneqq H(m_t|m_*)$ and $I_t \coloneqq I(m_t|m_*)$. Otto and Villani calculated that

$$\frac{\mathrm{d}H_t}{\mathrm{d}t} = -I_t,$$
$$\frac{\mathrm{d}I_t}{\mathrm{d}t} \leqslant -2\rho I_t.$$

Since we know that $\lim_{t\to\infty} H_t = 0$, we have

$$H_0 = \int_0^\infty I_t \,\mathrm{d}t \leqslant \int_0^\infty I_0 e^{-2\rho t}\,\mathrm{d}t = \frac{I_0}{2\rho}.$$

As the initial value of the flow is arbitrary, we have established the log-Sobolev inequality, and this leads to the exponential convergence of the relative entropy. So in the entropic fictitious play, the free energy $\mathcal{F}$ takes the role of entropy in Otto–Villani, and $H(m_t|\hat{m}_t)$ the role of Fisher information.

Despite the simplicity of the entropic fictitious play, an important numerical difficulty is not taken into account in the analysis above. At each step t, we need to compute the best response to m_t, namely $\hat{m}_t \propto \exp\bigl(-\frac{\delta F}{\delta m}(m_t,\cdot)\bigr)$, and this is usually done by Monte Carlo methods: for example, we launch particles from an initial distribution and let them evolve according to the overdamped Langevin dynamics. In sufficiently long time, with sufficiently large number of particles, we can sample the measure $\hat{m}_t$ with arbitrary precision. This step is called the inner iteration in Chapter 6. However, the computational complexity of this iteration is not addressed.

This is the reason in Chapter 7 we turn to the following dynamics:

$$\begin{aligned} \mathrm{d}X_t &= -D_m F(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \\ \mathrm{d}m_t &= \lambda(t)(\delta_{X_t} - m_t)\,\mathrm{d}t. \end{aligned} \tag{21}$$

Here $\lambda \colon [0,\infty) \to (0,\infty)$ is to be determined and the m_t is no longer the law of the particle X_t, but is a weighted occupation measure of the particle according to the second equation:

$$m_t = e^{-\int_0^t \lambda(s)\,\mathrm{d}s} m_0 + \int_0^t \lambda(s) e^{-\int_s^t \lambda(u)\,\mathrm{d}u}\,\delta_{X_s}\,\mathrm{d}s.$$

The drift of the particle at time t depends thus on its history on the interval $[0,t]$ and for this reason the dynamics is called *self-interacting*. This type of dynamics was already studied by Cranston, Le Jan [66], Raimond [188], and Benaïm, Ledoux and Raimond [15]. The recent paper of Du, Jiang and J. Li [75] addresses the utility of such dynamics in sampling. For now we fix $\lambda(t) = \lambda > 0$. Note that each of the two components in (21) has a natural time scale. If the measure argument m_t is frozen, the first component follows a linear overdamped Langevin and the time scale is the mixing time for such process. Similarly, by fixing the X_t argument in the second equation, we find that the time scale of the second component is $1/\lambda$. Under the limit $\lambda \to 0$, the second time scale becomes much larger than the first, so the distribution of the first argument relaxes quickly to the steady state $\hat{m}_t$ before the second argument changes significantly. And since, by Birkhoff's theorem, the Dirac mass δ_{X_t} averaged over a long enough interval is close to the steady state $\hat{m}_t$, we expect that in the long time the self-interacting dynamics should be effectively described by the entropic fictitious play:

$$\mathrm{d}m_t = \lambda(\hat{m}_t - m_t)\,\mathrm{d}t,$$

which converges to m_* when $t \to \infty$. On the other hand, under the limit $\lambda \to \infty$, the second argument m_t becomes very close to the Dirac mass δ_{X_t}, so the dynamics should be approximately the linear dynamics

$$\mathrm{d}X_t = -D_m F(\delta_{X_t}, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

This Markov process relaxes rapidly but its equilibrium measure, being proportional to $\exp\bigl(-F(\delta_x)\bigr)\,\mathrm{d}x$, is a priori different from our target m_*. We are thus in a situation similar to the classical bias-variance tradeoff once we make the connection between the relaxation rate and the inverse of variance.

We study this tradeoff quantitatively in Chapter 7. We take a fixed $\lambda > 0$ in the following and first study the convergence rate of the self-interacting process (21). Note that in this case, the process is a time-homogeneous Markov in an infinite-dimensional state space with a highly degenerate noise, so proving its ergodicity is usually a non-trivial task. However, thanks to the strong contractivity in the second argument, we are able to establish an explicit exponential contractivity in Wasserstein distance by a mixed coupling similar to the kinetic coupling of Eberle, Guillin and Zimmer [84]. The resulting contraction rate deteriorates as λ decreases, which is to be expected. Notably, the structural condition that F is flat-convex is not at all used for the Markov process's relaxation. As a by-product, we know that the stationary measure of the Markov process (21) exists and is unique, which we denote by P^λ.

Then we study the bias between the stationary measure P^λ and the target $m_* \otimes \delta_{m_*}$, where as we recall again, m_* is the invariant measure to the mean field Langevin dynamics (9) or the solution to the mean field optimization problem (11). To proceed, we suppose that the mean field dependency is cylindrical:

$$F(m) = \Phi(\langle \ell, m\rangle) = \Phi\biggl(\int_{\mathcal{X}} \ell(x) m(\mathrm{d}x)\biggr)$$

for some $\ell \colon \mathcal{X} \to \mathbb{R}^D$ and some convex $\Phi \colon \mathbb{R}^D \to \mathbb{R}$. Here the convexity of Φ implies the flat convexity of F as a mean field functional. Then the self-interacting process (21) can be reduced to the projected system:

$$\begin{aligned} \mathrm{d}X_t &= -\nabla_x V(Y_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \\ \mathrm{d}Y_t &= \lambda\bigl(\ell(X_t) - Y_t\bigr)\,\mathrm{d}t, \end{aligned} \tag{22}$$

where we have the correspondence

$$\begin{aligned} Y_t &= \langle \ell, m_t \rangle, \\ V(y,x) &= \nabla\Phi(y) \cdot \ell(x). \end{aligned}$$

Denote by $\rho \coloneqq \rho^\lambda$ the push-out of P^λ under the mapping

$$(x, m) \mapsto (x, \langle \ell, m \rangle).$$

By construction, the measure ρ is invariant to the reduced dynamics (22), and solves the stationary equation

$$\Delta_x \rho + \nabla_x \cdot \big(\nabla_x V(y,x)\rho\big) - \lambda \nabla_y \cdot \Big(\big(\ell(x) - y\big)\rho\Big) = 0.$$

Using the equation above and a uniform LSI, we derive the following L^1 estimate on conditional entropy:

$$\int_{\mathbb{R}^D} H\big(\rho^{1|2}(\cdot|y)\big|\hat{m}_y\big)\rho^2(\mathrm{d}y) = O(\lambda), \tag{23}$$

where $\rho^{1|2}$, ρ^2 are respectively the conditional and marginal measures formally defined by

$$\begin{aligned} \rho^2(y) &\coloneqq \int_{\mathcal{X}} \rho(x,y)\,\mathrm{d}x, \\ \rho^{1|2}(x|y) &\coloneqq \frac{\rho(x,y)}{\rho^2(y)}, \end{aligned}$$

and $\hat{m}_y$ is the Gibbs measure with density

$$\hat{m}_y(x) \propto \exp\big(-V(y,x)\big).$$

The estimate (23) indicates that on average, $\rho^{1|2}(\cdot|y)$ is close to $\hat{m}_y$. Denote the cylindrical projection of the target measure $y_* \coloneqq \langle \ell, m_* \rangle$. We notice that

$$\begin{aligned} &\int_{\mathbb{R}^D} \big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y) \\ &\qquad = -\int_{\mathcal{X}\times\mathbb{R}^d} \big(V(y,x) - V(y_*,x)\big)(\hat{m}_y - m_*)(\mathrm{d}x)\rho^2(\mathrm{d}y) \\ &\qquad = -\int_{\mathcal{X}\times\mathbb{R}^d} \big(V(y,x) - V(y_*,x)\big)\big(\rho^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y) + O(\sqrt{\lambda}), \end{aligned}$$

where for the last equality, we make the change of measure $\hat{m}_y \to \rho^{1|2}(\cdot|y)$ and control the error by the entropy estimate (23) and a transport inequality (Talagrand, Pinsker or Bolley–Villani [25] depending on the assumption on V). Using the form of the potential $V(y,x) = \nabla\Phi(y) \cdot \ell(x)$ and the convexity of Φ, we can show that

$$\int_{\mathcal{X}\times\mathbb{R}^d} \big(V(y,x) - V(y_*,x)\big)\big(\rho^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y) \geqslant 0.$$

So we get

$$\int_{\mathbb{R}^D} \big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y) = O(\sqrt{\lambda}).$$

Again using Talagrand's inequality, we find

$$\int_{\mathbb{R}^D} W_1\big(\rho^{1|2}(\cdot|y), m_*\big)\rho^2(\mathrm{d}y) \leqslant \int_{\mathbb{R}^D} \Big(W_1\big(\rho^{1|2}(\cdot|y), \hat{m}_y\big) + W_1(\hat{m}_y, m_*)\Big)\rho^2(\mathrm{d}y) = O(\lambda^{1/4}).$$

This already indicates that the measures P^λ and $m_* \otimes \delta_{m_*}$, projected into the X directions, are close to each other when λ is small. We can exploit again the gradient structure of the dynamics to show the same thing for the Y directions. Moreover, the order in λ can be improved to $O(\sqrt{\lambda})$. The final bound on the bias that we obtain is the following:

$$W(P^\lambda, m_* \otimes \delta_{m_*}) = O(\sqrt{\lambda}),$$

where W denotes a Wasserstein distance between finite-dimensional projections of the infinite-dimensional measures. This bound is also optimal in the order of λ, as it can be verified by a Gaussian example.

To summarize, a smaller λ leads to a weaker convergence rate, but reduces the bias of the sampling, confirming the intuitions from our previous discussions. However, it should be noted that the convergence rate achieved by reflection coupling deteriorates exponentially as $\lambda \to 0$, rendering this rate unsuitable for analyzing annealing dynamics in practice.

Preview of Chapter 8

Mean field Schrödinger dynamics

In the last chapter of the thesis, we study the mean field optimization problem regularized by Fisher information (12) and the associated gradient flow. As we mentioned above, the gradient flow should at least be the formal continuous limit of the discrete JKO scheme (13). By calculus of variation, we get that the discrete flow is in fact the backward Euler:

$$\frac{\delta \mathfrak{F}}{\delta m}(\nu^h_{n+1}, \cdot) + \frac{1}{h}\log\frac{\nu^h_{n+1}}{\nu^h_n} = \text{constant},$$

and we expect that $\nu^h_{\lfloor t/h \rfloor}$ converges to the flow m_t solving

$$\partial_t m_t = -\frac{\delta \mathfrak{F}}{\delta m}(m_t, \cdot)m_t + \lambda_t m_t,$$

where λ_t is the normalization constant

$$\lambda_t \coloneqq \int_{\mathcal{X}} \frac{\delta \mathfrak{F}}{\delta m}(m_t, x) m_t(\mathrm{d}x)$$

ensuring that the mass is conserved: $\mathrm{d}\int_{\mathcal{X}} m_t / \mathrm{d}t = 0$. Recall that the functional $\mathfrak{F}$ is regularized by Fisher information:

$$\mathfrak{F}(m) = F(m) + \int_{\mathcal{X}} \frac{|\nabla m|^2}{m}.$$

By integration by parts, we get the following expression for its linear functional derivative:
$$\frac{\delta \mathfrak{F}}{\delta m}(m,x) = \frac{\delta F}{\delta m}(m,x) - 2\nabla\cdot\left(\frac{\nabla m}{m}\right) - \frac{|\nabla m|^2}{m^2}.$$
We have also the gradient descent formula:
$$\frac{\mathrm{d}\mathfrak{F}(m_t)}{\mathrm{d}t} = -\int_{\mathcal{X}} \left|\frac{\delta \mathfrak{F}}{\delta m}(m_t,x) - \lambda_t\right|^2 m_t(\mathrm{d}x).$$
At this point, we can already expect that the non-linear flow m_t, once well defined, converges to the solution of optimization problem (12) for the following two reasons. First, by the formula above, the regularized energy $\mathfrak{F}$ ceases to decrease only if $\frac{\delta\mathfrak{F}}{\delta m}(m_t,\cdot) - \lambda_t = 0$, that is to say, the measure m_t is a stationary point to the optimization problem (12). Second, the optimization problem, being the sum of a flat-convex F (which is our standing structural condition) and a strictly flat-convex Fisher information, has only one stationary point, and this point is the global minimizer. Given these intuitions, we can rigorously prove the convergence by compactness and LaSalle's invariance principle, as done in [117].

The remaining question is to find an explicit convergence rate and the functional inequality behind this rate. In the case of overdamped mean field Langevin, the answer is a uniform log-Sobolev inequality as shown by [178, 56]. And for our mean field Fisher gradient flow, we need a uniform spectral gap, or in other words, a uniform Poincaré inequality. To see this, we make the change of variable
$$\psi_t \coloneqq \sqrt{m_t}$$
and write the dynamical equation for ψ_t:
$$\partial_t\psi_t = 2\Delta\psi_t - \frac{1}{2}\frac{\delta F}{\delta m}(\psi_t^2,\cdot)\psi_t + \frac{1}{2}\lambda_t\psi_t.$$
Now λ_t satisfies
$$\lambda_t = \int_{\mathcal{X}} 4|\nabla\psi_t|^2 + \frac{\delta F}{\delta m}(\psi_t,\cdot)\psi_t^2$$
and is the constant ensuring that ψ_t is normalized in L^2. In the linear case, the flat derivative does not depend on the measure:
$$\frac{\delta F}{\delta m}(m,x) = U(x),$$
and the evolution of ψ_t corresponds to a linear Schrödinger semigroup. The exponential convergence is thus guaranteed by the spectral gap of the Hamiltonian operator:
$$\mathcal{H} = -4\Delta + U.$$
Going back to our non-linear evolution, we define the Hamiltonian at each instant:
$$\mathcal{H}_t \coloneqq -4\Delta + \frac{\delta F}{\delta m}(m_t,\cdot),$$
we then have
$$\partial_t\psi_t = -\frac{1}{2}(\mathcal{H}_t - \lambda_t)\psi_t,$$
$$\lambda_t = (\psi_t, \mathcal{H}_t\psi_t)_{L^2}.$$

The descent of $\mathfrak{F}(m_t)$ satisfies

$$\frac{\mathrm{d}\mathfrak{F}(m_t)}{\mathrm{d}t} = -\big(\psi_t, (\mathcal{H}_t - \lambda_t)\mathcal{H}_t\psi_t\big)_{L^2} = -(\psi_t, \mathcal{H}_t^2\psi_t)_{L^2} + (\psi_t, \mathcal{H}_t\psi_t)_{L^2}^2.$$

Denote by $\hat\psi_t$ the unique normalized ground state of $\mathcal{H}_t$. We get, by the spectral gap,

$$(\psi_t, \mathcal{H}_t^2\psi_t)_{L^2} - (\psi_t, \mathcal{H}_t\psi_t)_{L^2}^2 \gtrsim (\psi_t, \mathcal{H}_t\psi_t)_{L^2} - (\hat\psi_t, \mathcal{H}_t\hat\psi_t)_{L^2}.$$

Again, going back to the measure variables and using the convexity of F, we can show that

$$(\psi_t, \mathcal{H}_t\psi_t)_{L^2} - (\hat\psi_t, \mathcal{H}_t\hat\psi_t)_{L^2} \geqslant \mathfrak{F}(m_t) - \inf\mathfrak{F}.$$

So we have the exponential convergence:

$$\mathfrak{F}(m_t) - \inf\mathfrak{F} \leqslant Ce^{-ct},$$

given the uniform spectral gap for $\mathcal{H}_t$. It is well known that the uniform spectral gap is equivalent to a uniform Poincaré inequality for the probability measure $\hat m_t \coloneqq \hat\psi_t^2$ solving the stationary equation

$$\frac{\delta F}{\delta m}(m_t, x) - 2\nabla\cdot\left(\frac{\nabla\hat m_t}{\hat m_t}\right) - \frac{|\nabla\hat m_t|^2}{\hat m_t^2} = \text{constant}.$$

Denoting the log-density by $\hat u_t \coloneqq -\log\hat m_t$, we find that $\hat u_t$ solves the ergodic HJB equation

$$2\Delta\hat u_t - |\nabla\hat u_t|^2 + \frac{\delta F}{\delta m}(m_t, x) = \text{constant}.$$

Under the assumption that $\frac{\delta F}{\delta m}(m,\cdot)$ is a sum of a strongly convex and a Lipschitz-continuous function, uniformly in m, we can employ the method of Conforti [61] to obtain that $\hat u_t$ is also a sum of a strongly convex and a Lipschitz part with uniform bounds. Then a uniform Poincaré inequality follows from for example [9].

Recent advances and perspectives

A common drawback of Chapters 1 and 2, as an anonymous referee has put it, is that we do *not directly* compare the particle system m_t^N and the mean field flow m_t in the long time. Instead, this comparison is done via the mean field invariant measure m_*, complemented with a standard global-in-time bound. This triangle argument is rather awkward and leads to loss of exponent in the final propagation of chaos bound. We announce that we will solve this problem by a direct comparison method, where we work with a distance between probability measures that is induced by the free energy landscape, and recover the optimal $O(1)$ (or $O(1/N)$, depending on the scaling) order error bound. We will also explore other consequences of the non-linear LSI (17) and its N-particle version (16), such as time-uniform measure concentration for the mean field Langevin particle system and turnpike properties for the associated mean field Schrödinger problem.

In a recent work by the author [230], the defective LSI (16) established in Chapter 1 was tightened to an N-uniform LSI through the use of an additional Poincaré inequality. This approach offers an alternative to the concurrent work of Chewi, Nitanda, and M. S. Zhang [55], while providing improved dependence on the mean field interaction strength.

More recently, Bauerschmidt, Bodineau and Dagallier [14] adapted the Polchinski flow method to mean field particle systems and established N-uniform LSI throughout the entire uniqueness regime. Specifically, the free energy functional $\mathcal{F}$ is permitted to include a *flat-concave* energy component, and the analysis is conducted directly under a projected form of the non-linear LSI:

$$I(m|\hat{m}) \gtrsim H(m|\hat{m}) \gtrsim \mathcal{F}(m) - \mathcal{F}(m_*).$$

This assumption is weaker than flat convexity and enables the recovery of Curie–Weiss critical behavior. However, the method appears less intrinsic for non-quadratic interactions and results in a weaker LSI constant. We announce here that, in a forthcoming work, we will establish an N-uniform defective LSI via an intrinsic approach that depends solely on the unprojected free energy landscape. As a further remark, our method corresponds to a *coordinate stochastic localization* scheme, whereas theirs is a *linear tilt* scheme in the language of Y. Chen–Eldan [52].

In another recent work [194], we investigate the size of chaos problem for the overdamped Langevin dynamics under the aforementioned non-linear LSI condition. More precisely, we show that

$$H(m_*^{N,k} \mid m_*^{\otimes k}) = O\left(\frac{k^2}{N^2}\right),$$

where $m_*^{N,k}$ denotes the k-marginal distribution of the N-particle Gibbs measure. In that work, we identify a gradient structure for conditional measures and develop an entropy hierarchy that is one order higher than Lacker's original formulation. This non-perturbative approach to mean field interaction extends the existing literature [141, 142, 20], which addresses only scenarios where the interaction is effectively dominated by diffusion. Nevertheless, the dynamical problem of uniform-in-time sharp chaos remains largely open and clearly warrants further investigation.

For the singular 2D vortex model, the size of chaos problem is not completely solved in the current thesis as our method fails in the low-temperature regime. The full resolution of this problem requires additional study but it seems to the author that some crucial elements are still lacking. Furthermore, we can also consider the size of chaos problem for Coulomb or Riesz interactions in higher dimension. This seems even more difficult to the author due to the increased singularity of the interaction kernel.

The study of Vlasov–Poisson systems has recently seen significant advances, with several novel ideas and techniques for establishing propagation of chaos introduced in [30, 29, 51]. However, the unregularized case in dimension $\geqslant 3$, for both the diffusive and non-diffusive settings, remains an open problem.

We may also ask whether the crucial elliptic entropy estimate in Chapter 7 can be extended to the dynamical parabolic case. If successful, such an approach would yield stronger contractivity properties than those achieved by the coupling method. We also intend to investigate kinetic self-interacting dynamics and explore the use of self-interaction in addressing mean field games.

As the studies of the mean field Schrödinger dynamics in Chapter 8 focus on the theoretical part, it is equally important to explore its numerical aspects and effectiveness in real-world applications.

*
* *

The eight chapters of the thesis have first appeared individually as the publications [50, 48, 171, 170, 49] and preprints [229, 77, 60]. For this reason, the notations and conventions in different chapters may not be consistent. They may also be different from these used in this introduction.

Introduction

L'objectif de cette thèse est d'étudier le comportement asymptotique de diverses dynamiques avec des *interactions de champ moyen*. Bien qu'il semble difficile de donner une définition à la fois précise et générale des interactions de champ moyen, nous pouvons dire que, dans le cadre de cette thèse, nous nous intéressons aux *flots non linéaires de mesures de probabilité* qui sont au moins des limites formelles de systèmes de particules en interaction, où l'équation de mouvement de chaque particule est influencée par les autres particules d'une manière plus ou moins égale et la force totale des influences est d'ordre 1. Pour donner un exemple concret, considérons les *dynamiques de McKean-Vlasov* décrites par l'équation différentielle stochastique (EDS) suivante :

$$\mathrm{d}X_t = b(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \text{où } m_t = \mathrm{Loi}(X_t). \tag{1}$$

Ici, la solution X_t de l'EDS est supposée exister sur la demi-droite $[0, \infty)$ et prendre des valeurs dans $\mathcal{X}$, où $\mathcal{X}$ est l'espace euclidien $\mathbb{R}^d$ ou le tore $\mathbb{T}^d$ pour un entier $d \geqslant 1$; la dérive $b\colon \mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}^d$ est suffisamment régulière pour assurer le caractère bien posé de l'équation ; et B est le mouvement brownien standard d-dimensionnel. Pour passer du point de vue probabiliste au point de vue analytique, on écrit l'équation aux dérivées partielles (EDP) d'évolution que m_t doit satisfaire :

$$\partial_t m_t = \Delta m_t - \nabla \cdot \big(b(m_t, \cdot) m\big). \tag{2}$$

La non-linéarité de l'équation ci-dessus est due à la dépendance de la mesure dans la dérive $b(\cdot, \cdot)$. On dit que la non-linéarité est de type champ moyen s'il existe une application

$$\frac{\delta b}{\delta m} : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$$

qui approxime la non-linéarité dans le sens que

$$b\big((1-t)m + t\delta_y, x\big) - b(m, x) = t\frac{\delta b}{\delta m}(m, x, y) - t\int_{\mathcal{X}} \frac{\delta b}{\delta m}(m, x, z) m(\mathrm{d}z) + o(t)$$

sous la limite $[0, 1] \ni t \to 0$, pour tout $m \in \mathcal{P}(\mathcal{X})$ et x, $y \in \mathcal{X}$, où δ_y représente la masse de Dirac en y. Cela exclut notamment les interactions locales où la dérive $b(m, x)$ dépend de la densité locale $m(x)$ de la mesure, c'est-à-dire $b(m, x) = \beta\big(m(x), x\big)$ pour une certaine fonction $\beta\colon \mathbb{R} \times \mathcal{X} \to \mathbb{R}^d$. Cela exclut aussi malheureusement le célèbre modèle de Boltzmann en théorie cinétique, où seules les particules à la même position spatiale sont autorisées à s'interagir.

Nous nous tournons maintenant vers le système de particules qui correspond à l'EDS (1) ou à l'EDP de Fokker-Planck (2). Soit N un entier $\geqslant 1$ et notons l'intervalle entier $[\![1,N]\!] = \{1,\dots,N\}$ par $[N]$. On introduit la notation abrégée $\boldsymbol{x} \coloneqq (x^1,\dots,x^N) \in \mathcal{X}^N$ pour le N-uplet d'éléments dans $\mathcal{X}$ et note la mesure empirique correspondante par

$$\mu^N_{\boldsymbol{x}} \coloneqq \frac{1}{N}\sum_{i\in[N]}\delta_{x^i}.$$

Le système d'EDS des N particules s'écrit

$$\mathrm{d}X^i_t = b(\mu^N_{\boldsymbol{X}_t}, X^i_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B^i_t, \qquad \text{pour } i \in [N], \tag{3}$$

où B^i sont des mouvements browniens standards indépendants en d dimensions et $\mu^N_{\boldsymbol{X}_t}$ est la mesure empirique formée par $\boldsymbol{X}_t = (X^1_t,\dots,X^N_t)$. Passant au côté analytique, on peut également écrire l'EDP d'évolution à N particules pour $m^N_t \coloneqq \mathrm{Loi}(\boldsymbol{X}_t)$:

$$\partial_t m^N_t = \sum_{i\in[N]}\Delta_i m^N_t - \sum_{i\in[N]}\nabla_i\cdot\big(b(\mu^N_{\boldsymbol{x}},x^i)m^N_t\big). \tag{4}$$

L'équation de Fokker-Planck ci-dessus est définie sur $[0,\infty)\times\mathcal{X}^N$ et est principalement une équation *linéaire*, car la dérive

$$b^{N,i}(\boldsymbol{x}) \coloneqq b(\mu^N_{\boldsymbol{x}}, x^i), \qquad \text{pour } i\in[N]$$

est entièrement déterminée par la configuration des particules $\boldsymbol{x} = (x^1,\dots,x^N)$ et ne dépend plus d'une loi de probabilité externe. En quelque sorte, on a éliminé la non-linéarité de la dynamique au prix d'augmenter considérablement la dimension de l'EDP.

Comme nous l'avons mentionné précédemment, on s'attend à ce que lorsque N tend vers l'infini, le système non linéaire décrit par (1) ou (2) fournisse une bonne approximation de la dynamique à N particules (3) ou (4). Plus précisément, on s'attend à ce que si les N particules sont initialisées indépendamment de m_0, c'est-à-dire

$$\mathrm{Loi}(\boldsymbol{X}_0) = m^N_0 = m_0^{\otimes N} = \mathrm{Loi}(X_0)^{\otimes N},$$

alors la limite

$$\frac{1}{N}\sum_{i\in[N]}\delta_{X^i_t} = \mu^N_{\boldsymbol{X}_t} \to m_t \text{ en probabilité}, \qquad \text{lorsque } N\to\infty \tag{5}$$

est vérifiée pour tout $t>0$. Il s'agit d'une *loi des grands nombres* pour les systèmes de particules en interaction. De plus, si les particules sont échangeables, c'est-à-dire que la loi conjointe des particules ne dépend pas de leur ordre, alors la convergence de la mesure empirique ci-dessus est équivalente à la convergence

$$\mathrm{Loi}(X^1_t,\dots,X^k_t) \eqqcolon m^{N,k}_t \to m_t^{\otimes k} \text{ faiblement, lorsque } N\to\infty, \text{ pour tout } k \text{ fixé}, \tag{6}$$

ou en d'autres termes, le sous-système de k particules est *asymptotiquement indépendant* quand la taille du système entier tend vers l'infini. Voir par exemple le lemme 1.1.2 de la thèse de Le Bris [145] pour un énoncé précis de cette équivalence. Pour des raisons historiques, le fait que les particules soient asymptotiquement indépendantes les unes des autres est appelé *chaos* dans les premières littératures en théorie cinétique, et donc la limite de champ moyen ci-dessus est appelée *propagation du chaos* : une fois que l'on a une condition initiale chaotique, alors les particules sont chaotiques en un temps positif.

Il a été découvert par Sznitman [216] au début des années 1990 que grâce à la technique de *couplage synchrone*, on peut prouver la propagation du chaos pour une grande classe de dérives. Nous expliquons les idées principales de cette méthode dans cet alinéa. La première étape de la méthode consiste à créer N copies indépendantes de l'EDS de champ moyen (1), ou en termes analytiques, à considérer la N-ième tensorisation $m_t^{\otimes N}$ du flot non linéaire $m_t = \mathrm{Loi}(X_t)$. On désigne par $\bar{X}^i$, $i \in [N]$, ces solutions indépendantes de l'EDS de champ moyen, et par $\bar{B}^i$, $i \in [N]$, les bruits browniens indépendants qui dirigent la dynamique. Ils satisfont donc les EDS suivantes :

$$\mathrm{d}\bar{X}_t^i = b(m_t, \bar{X}_t^i)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\bar{B}_t^i, \qquad \text{pour } i \in [N].$$

Ensuite, on couple les dynamiques indépendantes $\bar{X}_t^i$ ci-dessus avec la dynamique en interaction (3) en demandant que les bruits browniens soient exactement les mêmes, ou en d'autres termes synchronisés :

$$B^i = \bar{B}^i, \quad \text{pour } i \in [N].$$

En soustrayant ensuite l'EDS des particules en interaction et des particules indépendantes, on obtient

$$\mathrm{d}(X_t^i - \bar{X}_t^i) = \big(b(\mu_{\mathbf{X}_t}^N, X_t^i) - b(m_t, \bar{X}_t^i)\big)\,\mathrm{d}t,$$

où les bruits sont complètement annulés. Supposons maintenant que le coefficient de dérive b est suffisamment régulier pour que l'on ait le contrôle suivant :

$$|b(\mu_{\mathbf{X}_t}^N, X_t^i) - b(m_t, \bar{X}_t^i)| \lesssim \frac{1}{N} \sum_{j\in[N]} |X_t^j - \bar{X}_t^j| + |X_t^i - \bar{X}_t^i| + R_t, \tag{7}$$

où R_t est une variable aléatoire positive telle que $\mathbb{E}[R_t] \to 0$ quand $N \to \infty$. Dans le travail original de Sznitman, la dérive dépend de la mesure à travers une fonction noyau :

$$b(m, x) = \int_{\mathcal{X}} K(x, y) m(\mathrm{d}y)$$

pour une certaine $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$. En travaillant avec un K suffisamment régulier, Sznitman a montré que le terme d'erreur R_t correspond à l'erreur entre

$$\frac{1}{N} \sum_{j\in[N]} K(\cdot, \bar{X}_t^j) \quad \text{et} \quad \int_{\mathcal{X}} K(\cdot, y) m_t(\mathrm{d}y).$$

Comme $\bar{X}_t^j$, $j \in [N]$, sont des variables indépendantes de loi m_t, le terme d'erreur R_t peut être contrôlé par $O(N^{-1/2})$ grâce à l'argument classique de la variance, ce qui est l'ordre optimal en N selon le théorème central limite. On peut également vérifier ce contrôle pour un b qui est lipschitzien conjointement en mesure et en espace, où la distance métrique pour l'argument de mesure est la *distance de Kantorovich* ou la *distance de Wasserstein de* L^1. Notamment, une récente percée de Fournier et Guillin [93] nous permet d'identifier l'ordre optimal en N (environ

$O(N^{-1/d})$) pour le terme d'erreur aléatoire R_t dans le cas de caractère lipschitzien de Wasserstein. Une fois que l'on établit le contrôle (7), en prenant les valeurs absolues et en sommant sur $i \in [N]$, on obtient

$$\mathrm{d} \sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] \lesssim \sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|]\,\mathrm{d}t + N\,\mathbb{E}[R_t]\,\mathrm{d}t.$$

Ensuite, selon le lemme de Grönwall, on obtient

$$\frac{1}{N}\sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] \leqslant \frac{e^{Ct}}{N}\sum_{i\in[N]} \mathbb{E}[|X_0^i - \bar{X}_0^i|] + C\int_0^t e^{C(t-s)}\,\mathbb{E}[R_s]\,\mathrm{d}s.$$

Dans le cas où la condition initiale est chaotique $m_0^N = m_0^{\otimes N}$, on peut prendre $\bar{X}_0^i = X_0^i$ de sorte que le premier terme du côté droit s'annule. En utilisant le fait que $\mathbb{E}[R_\cdot] \to 0$ lorsque $N \to \infty$, on obtient

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] = 0$$

ce qui suffit à justifier la limite de champ moyen $\mu^N_{\mathbf{X}_t} \to m_t$ dans (5).

La méthode de couplage synchrone présentée ci-dessus, bien que simple et robuste, échoue à aborder le *comportement en temps long* de la limite de champ moyen sans des conditions supplémentaires. En effet, dans le cas où le terme d'erreur satisfait la borne uniforme

$$\mathbb{E}[R_s] \leqslant \frac{C}{\sqrt{N}},$$

on peut seulement obtenir, en modifiant la constante C,

$$\frac{1}{N}\sum_{i\in[N]} \mathbb{E}[|X_t^i - \bar{X}_t^i|] \leqslant \frac{C(e^{Ct}-1)}{\sqrt{N}}.$$

Cela signifie que l'on a besoin d'un nombre exponentiellement grand de particules pour bien approximer le flot de champ moyen non linéaire sur le long terme. Ce phénomène est générique dans les systèmes évolutifs (rappelons-nous la théorie de Cauchy-Lipschitz pour les EDO) et on doit imposer des conditions structurelles pour éviter de telles croissances exponentielles de l'erreur en temps.

Flots de gradient et convexités

La principale condition structurelle sur la dynamique dans la thèse est que la dérive est un gradient de Wasserstein négatif correspondant à un problème d'*optimisation de champ moyen* convexe. Pour être précis, soit $F\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ une fonctionnelle de champ moyen. On dit que F admet un gradient de Wasserstein $D_mF\colon \mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}^d$ si

$$\lim_{t\searrow 0} \frac{F\big((e^{tv})_\# m\big) - F(m)}{t} = \int_{\mathcal{X}} D_mF(m,x)\cdot v(x)m(\mathrm{d}x)$$

pour toute mesure $m \in \mathcal{P}(\mathcal{X})$ et tout champ de vecteurs $v\colon \mathcal{X} \to \mathbb{R}^d$ suffisamment régulier. Ici, e^{tv} désigne l'application exponentielle engendrée par le champ de vecteurs qui correspond à l'EDO $\dot{x} = v(x)$ de la manière suivante :

$$e^{tv}x_0 = x_t, \qquad \text{où } x\colon [0,t] \to \mathcal{X} \text{ résout } \dot{x}_s = v(x_s) \text{ pour } s \in [0,t]\,;$$

et $(e^{tv})_\# m$ désigne l'image de la mesure m par l'application e^{tv}. La condition structurelle que l'on a imposée ci-dessus peut être précisément formulée comme suit :

$$b(m,x) = -D_m F(m,x) \qquad \text{pour une convexe } F\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}. \tag{8}$$

Ici, la convexité de F est comprise dans le sens d'interpolation plate suivant :

$$\forall m_0, m_1 \in \mathcal{P}(\mathcal{X}),\ \forall t \in [0,1], \qquad F\big((1-t)m_0 + tm_1\big) \leqslant (1-t)F(m_0) + tF(m_1),$$

et cela ne doit pas être confondu avec la convexité de déplacement, où l'on construit l'interpolation entre les mesures de probabilité par le transport optimal (voir le chapitre 1 pour plus de discussions sur la différence entre les deux notions de convexité). Pour clarifier les idées, supposons que la fonctionnelle de champ moyen F satisfait

$$F(m) = \int_{\mathcal{X}} U(x) m(\mathrm{d}x)$$

pour une fonction de potentiel $U\colon \mathcal{X} \to \mathbb{R}$ suffisamment régulière. Alors le gradient de Wasserstein de F est simplement ∇U (qui ne dépend pas de la mesure) et F est toujours linéaire (donc convexe) dans le sens d'interpolation plate. Mais F est convexe de déplacement si et seulement si le potentiel sous-jacent U est une fonction convexe (voir les discussions dans [4, Chapter 9]). Dans ce cas, les EDS (1), (3) deviennent

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

qui est la *dynamique de Langevin suramortie* classique. Ainsi, la dynamique de champ moyen de notre intérêt

$$\mathrm{d}X_t = -D_m F(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \tag{9}$$

est appelée *dynamique de Langevin de champ moyen suramortie* et de manière similaire pour son système correspondant de particules. Passant au côté analytique, on peut écrire l'EDP associée

$$\partial_t m_t = \Delta m_t + \nabla \cdot \big(D_m F(m_t, \cdot) m_t\big). \tag{10}$$

On mentionne que le gradient de Wasserstein d'une fonctionnelle de champ moyen est également lié à sa dérivée fonctionnelle linéaire, dont la définition précise sera donnée dans les chapitres suivants. La dérivée linéaire est notée $\delta F/\delta m$ et est une application de $\mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}$. Sous une régularité suffisante, ces deux dérivées satisfont l'égalité suivante :

$$D_m F(m,x) = \nabla_x \frac{\delta F}{\delta m}(m,x).$$

La raison pour laquelle la condition (8) mène aux propriétés en temps long pour le flot de champ moyen (1) est due à une observation simple mais puissante de Jordan, Kinderlehrer et Otto [126] à la fin des années 1990 : le flot des mesures associé à l'EDS (9) est un *flot de gradient* pour la fonctionnelle d'*énergie libre*

$$\mathcal{F}(m) = F(m) + H(m), \qquad \text{où } H(m) = \int_{\mathcal{X}} m(x) \log m(x)\,\mathrm{d}x$$

dans l'espace de Wasserstein de L^2. En particulier, le long du flot $t \mapsto m_t$, l'énergie libre $t \mapsto \mathcal{F}(m_t)$ décroît. Comme la convexité de F assure que l'énergie libre $\mathcal{F} =$

$F + H$ a un minimiseur unique m_*, on peut s'attendre à ce que le flot de champ moyen converge vers m_*. En d'autres termes, le flot (10) fournit une méthode dynamique pour résoudre le problème d'optimisation *régularisé par l'entropie* :

$$\inf_{m \in \mathcal{P}(\mathcal{X})} \mathcal{F}(m) = \inf_{m \in \mathcal{P}(\mathcal{X})} F(m) + H(m). \tag{11}$$

Plus précisément, en notant W_2 la distance de Wasserstein de L^2 et h le pas de temps > 0, on peut définir de manière itérative le flot discret suivant de mesures de probabilité :

$$\mu_{n+1}^h = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \mathcal{F}(\mu) + \frac{W_2^2(\mu, \mu_n^h)}{2h}, \qquad \text{avec la condition initiale } \mu_0^h = m_0.$$

Ce schéma discret est appelé *schéma de JKO*. Et l'on a la limite

$$\mu_{\lfloor t/h \rfloor}^h \to m_t, \qquad \text{lorsque } h \to 0, \qquad \text{pour tout } t > 0.$$

Le lecteur peut se convaincre de la terminologie de « flot de gradient » en considérant l'analogie suivante en dimension finie. Soit n un entier $\geqslant 1$ et $V \colon \mathbb{R}^n \to \mathbb{R}$ une fonction potentielle. La dynamique discrète définie par

$$x_{n+1}^h = \operatorname*{argmin}_{x \in \mathbb{R}^n} V(x) + \frac{|x - x_n^h|^2}{2h}, \quad \text{avec la condition initiale } x_0^h = x_0$$

n'est autre que le schéma d'Euler implicite

$$x_{n+1}^h = x_n^h - h\nabla V(x_{n+1}^h)$$

pour la descente de gradient

$$\dot{x} = -\nabla V(x),$$

et sous des hypothèses de régularité sur V, on peut montrer que la dynamique discrète converge vers l'EDO continue. Travaillant avec la convexité de déplacement, Carrillo, McCann et Villani [39] ont étudié la dissipation de l'énergie libre et ont obtenu l'ergodicité du flot non linéaire (9) dans les années 2000. Ambrosio, Gigli et Savaré ont ensuite traduit beaucoup de résultats obtenus sous la convexité de déplacement en énoncés dans le formalisme abstrait des flots de gradients dans les espaces métriques, magnifiquement présentés dans leur monographie [4]. D'autre part, seulement récemment la structure de flot de gradient et la convexité plate ont été exploitées pour obtenir les comportements en temps long du flot de Langevin de champ moyen suramorti. Nous mentionnons ici les travaux de K. Hu, Z. Ren, Šiška et Szpruch [117], de Nitanda, D. Wu et Suzuki [178], et de Chizat [56].

Les motivations derrière nos études sur la convexité plate pour la Langevin de champ moyen ou le problème d'optimisation de champ moyen sont à la fois théoriques et pratiques. D'un point de vue théorique, il est naturel de chercher à dépasser les littératures classiques qui reposent sur la convexité de déplacement et d'explorer les convexités alternatives conduisant aux comportements en temps longs. La convexité plate est l'un des candidats naturels. En fait, de manière intéressante, pour les systèmes de jeux à champ moyen (JCM), qui consistent essentiellement en une paire d'équations de Fokker-Planck et de Hamilton-Jacobi-Bellman couplées entre elles, la condition classique assurant le caractère bien posé sur des intervalles arbitrairement longs est la monotonie de Lasry-Lions [143], ou la convexité plate dans le cas des jeux de potentiel. Un peu plus tard, Gangbo et Mészáros [94] ont montré que la convexité de déplacement est suffisante pour le caractère bien posé global du JCM. Pour la partie pratique, il y a récemment un intérêt croissant à modéliser les dynamiques d'entraînement des réseaux neuronaux comme un flot de gradient dans l'espace de mesures de probabilité, et dans le cas des réseaux peu profonds, le paysage de perte est convexe au sens plat [163, 57, 211, 203]. Le lecteur peut se référer aux sections d'application des chapitres 1 et 2 pour une introduction détaillée aux réseaux neuronaux peu profonds et leur formulation de champ moyen.

Principales contributions

L'une des contributions majeures de cette thèse est d'étudier non seulement le flot de champ moyen (9) à long terme, mais aussi le système de particules associé, sous la convexité plate du fonctionnel d'énergie F. Du point de vue numérique, c'est la question naturelle à poser après avoir établi les comportements à long terme de la limite de champ moyen. En effet, pour l'exemple de réseau de neurones mentionné ci-dessus, le flot de champ moyen m_t correspond à la dynamique d'apprentissage d'un réseau à un nombre infini de neurones et n'est pas accessible aux ordinateurs réels. La véritable dynamique d'apprentissage implique toujours un nombre fini de particules et n'est que des approximations de la limite de champ moyen. En rappelant que m_t^N est la loi conjointe des N particules, nous souhaitons montrer qu'il existe une borne uniforme dans le temps sur l'erreur d'approximation

$$d(m_t^N, m_t^{\otimes N})$$

pour une métrique appropriée d sur l'espace de mesures de probabilité. Cette propriété, appelée *propagation du chaos uniforme en temps*, est l'objectif principal des deux premiers chapitres, qui forment la partie I de la thèse. Dans le chapitre 1, nous développons les idées expliquées ci-dessus et montrons que l'erreur d'approximation est uniformément bornée dans le temps. Dans le chapitre 2, nous étudions la variante cinétique des dynamiques de Langevin de champ moyen et obtenons des résultats similaires.

Un autre ingrédient clé de la partie I est l'inégalité de Sobolev logarithmique (inégalité de log-Sobolev, ISL), et nous en illustrons l'importance en résumant l'argument de [178, 56] comme suit. Comme nous l'avons mentionné précédemment, la méthode de la partie I repose sur la structure du gradient et la dissipation d'énergie libre associées à l'équation de Fokker-Planck non linéaire (10). En prenant la dérivée temporelle de la fonctionnelle d'énergie libre, on obtient, au moins formellement,

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\int_{\mathcal{X}} |\nabla \log m_t + D_m F(m_t, \cdot)|^2 \, \mathrm{d}m_t.$$

On définit $\hat{m}$ comme l'unique mesure de probabilité qui a pour densité

$$\hat{m}(x) \propto \exp\biggl(-\frac{\delta F}{\delta m}(m, x)\biggr),$$

où $\delta F/\delta m$ est la dérivée fonctionnelle linéaire de F. En utilisant ensuite la relation entre la dérivée linéaire et le gradient de Wasserstein, on trouve

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\int_{\mathcal{X}} \left|\nabla \log \frac{m_t}{\hat{m}_t}\right|^2 \mathrm{d}m_t =: -I(m_t|\hat{m}_t),$$

où la fonctionnelle $I(\cdot|\cdot)$ est appelée *information de Fisher relative*. L'inégalité de Sobolev logarithmique pour la mesure $\hat{m}_t$ nous permet alors de minorer l'information de Fisher $I(m_t|\hat{m}_t)$ par l'entropie relative

$$H(m_t|\hat{m}_t) \coloneqq \int_{\mathcal{X}} \log \frac{m_t(x)}{\hat{m}_t(x)} m_t(\mathrm{d}x),$$

à une constante multiplicative près. Grâce à la convexité plate, l'entropie relative $H(m_t|\hat{m}_t)$ peut à nouveau être minorée par l'énergie libre relative $\mathcal{F}(m_t) - \inf \mathcal{F}$. Ainsi, en combinant l'ISL et la convexité, on démontre la contractivité exponentielle de l'énergie libre.

Étant un outil puissant pour obtenir la contractivité exponentielle, les inégalités de Sobolev logarithmiques sont malheureusement difficiles à établir, surtout lorsque l'on a pas d'accès direct à la densité de la mesure de probabilité concernée. Et c'est là l'objectif principal des deux premiers chapitres de la partie II. Dans les chapitres 3 et 4, nous proposons deux classes de critères pour l'inégalité de Sobolev logarithmique basées sur deux méthodes complètement différentes. Nous donnons ensuite quelques applications de l'ISL aux comportements en temps longs des systèmes de particules dans la seconde moitié du chapitre 4. Notamment, nous étendons la propagation du chaos locale uniforme en temps de Lacker et Le Flem [142] au cas d'un potentiel d'interaction non convexe. En approfondissant l'une des méthodes plus en détail, nous parvenons à démontrer une borne de L^∞ sur le hessien de la densité logarithmique du flot de champ moyen (2) avec le noyau de Biot-Savart (contenant une singularité) défini sur l'*espace entier*. Ce résultat technique nous permet de démontrer pour la première fois la propagation du chaos uniforme en temps pour le modèle de vortex en 2D sur l'espace entier. En tant que suite du chapitre 4, nous travaillons toujours sur le modèle de vortex en 2D dans le chapitre 5 et montrons la propagation du chaos uniforme et optimale dans le régime de haute température en étendant la méthode de Lacker [140] aux interactions singulières. La principale nouveauté de notre méthode est l'utilisation d'une technique combinatoire pour résoudre la hiérarchie des entropies qui implique des termes supplémentaires introduits par l'interaction singulière.

Enfin, nous abordons la dernière et peut-être la partie la plus exotique de la thèse, où nous étudions les comportements en temps long des dynamiques non linéaires en dehors du cadre de McKean-Vlasov qui a été discuté jusqu'à présent. Néanmoins, dans les chapitres 6 et 7, nous nous concentrons toujours sur le problème d'optimisation de champ moyen régularisé par l'entropie (11). Remarquons que la condition du premier ordre du problème est équivalente au problème du point fixe

$$m = \hat{m},$$

où, comme nous nous en souvenons, $\hat{m}(x) \propto \exp\bigl(-\frac{\delta F}{\delta m}(m,x)\bigr)$. En interprétant $\hat{m}$ comme la meilleure réponse à m, la formulation du point fixe peut être comprise comme une condition d'équilibre de Nash pour un auto-jeu où une personne joue contre elle-même. Motivés par la stratégie de jeu fictif de la théorie classique des jeux, nous étudions la dynamique suivante dans le chapitre 6 :

$$\partial_t m_t = \alpha(\hat{m}_t - m_t),$$

pour une constante $\alpha > 0$, appelée *jeu fictif entropique*, et démontrons sa convergence vers l'équilibre. Un inconvénient majeur du jeu fictif entropique est que, donné l'état d'un joueur m, il peut être coûteux de calculer la meilleure réponse $\hat{m}$, car cela nécessite généralement un calcul de type Monte-Carlo. Pour surmonter ce problème, dans le chapitre 7, nous proposons une dynamique de diffusion auto-interagissante qui peut être considérée comme un régime intermédiaire entre le jeu fictif entropique et un processus de diffusion linéaire. Nous expliquerons ce point en détail ci-dessous. La convergence à l'équilibre pour cette dynamique auto-interagissante est également établie dans ce chapitre. Dans le dernier chapitre 8, nous nous intéressons plutôt au problème d'optimisation de champ moyen *régularisé par l'information de Fisher* :

$$\inf_{m\in\mathcal{P}(\mathcal{X})} \mathfrak{F}(m) \coloneqq \inf_{m\in\mathcal{P}(\mathcal{X})} F(m) + I(m) \coloneqq \inf_{m\in\mathcal{P}(\mathcal{X})} F(m) + \int_{\mathcal{X}} \frac{|\nabla m(x)|^2}{m(x)}\,\mathrm{d}x, \tag{12}$$

et la descente de gradient associée, avec l'entropie relative mesurant la distance entre les mesures de probabilité. En d'autres termes, nous proposons d'étudier la limite continue du schéma de JKO suivant :

$$\nu^h_{n+1} \coloneqq \operatorname*{argmin}_{\nu\in\mathcal{P}(\mathcal{X})} \mathfrak{F}(\nu) + \frac{H(\nu|\nu^h_n)}{h}, \qquad \text{avec la condition initiale } \nu^h_n = m_0. \tag{13}$$

La dynamique résultante est une version non linéaire du semigroupe de Schrödinger et est donc appelée *dynamique de Schrödinger de champ moyen*. On obtient sa convergence exponentielle via un trou spectral uniforme, c'est-à-dire une inégalité de Poincaré uniforme.

*
* *

Dans le reste de cette introduction, nous présentons des aperçus techniques détaillés des huit chapitres de la thèse. Nous discutons également quelques perspectives à la fin.

Aperçu des chapitres 1 et 2

Propagation du chaos uniforme pour les Langevin

Les résultats principaux des deux chapitres sont la propagation du chaos uniforme en temps pour les dynamiques de Langevin de champ moyen suramortie et sousamortie. Comme la dynamique suramortie a été définie précédemment dans (9), on définit ici uniquement la dynamique sousamortie, ou cinétique,

$$\begin{aligned} \mathrm{d}X_t &= V_t\,\mathrm{d}t, \\ \mathrm{d}V_t &= -V_t\,\mathrm{d}t - D_m F(m^X_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \text{où } m^X_t = \mathrm{Loi}(X_t). \end{aligned} \tag{14}$$

La structure du seconde ordre de la dynamique modélise une particule newtonienne soumise à des forces aléatoires, ce qui la rend plus apte à décrire les phénomènes physiques. En outre, la dynamique de Langevin cinétique présente un analogue de l'accélération de Nesterov pour les méthodes de Monte-Carlo par chaînes de Markov de type gradient, c'est-à-dire la Langevin suramortie. Voir le travail de Y.-A. Ma et al. [157]. Le système de N particules associé est défini en dupliquant l'EDS N fois et en remplaçant la dépendance de m^X_t par la mesure empirique

$$\mu^N_{\mathbf{X}_t} = \frac{1}{N} \sum_{i\in[N]} \delta_{X^i_t}.$$

Nous notons toujours par m_t^N la loi conjointe des N particules, mais remarquons que maintenant cette loi est conjointe à la fois en espace et en vitesse :

$$m_t^N \coloneqq \mathrm{Loi}\big((X_t^1, V_t^1), \dots, (X_t^N, V_t^N)\big).$$

Notre approche des comportements en temps long des dynamiques de Langevin de champ moyen suramortie et sousamortie se base sur l'(hypo-)coercivité entropique de la dynamique, que nous expliquons en détail par la suite.

Considérons pour l'instant le cas suramorti et désignons par m_* la mesure invariante unique de (2). Cette mesure est également le minimiseur unique de la fonctionnelle d'énergie libre de champ moyen :

$$m_* = \operatorname*{argmin}_{m \in \mathcal{P}(\mathcal{X})} \mathcal{F}(m) = \operatorname*{argmin}_{m \in \mathcal{P}(\mathcal{X})} F(m) + H(m).$$

Introduisons la fonctionnelle d'énergie libre relative :

$$\begin{aligned} \mathcal{F}^N(m_t^N | m_*) &\coloneqq \mathcal{F}^N(m_t^N) - N\mathcal{F}(m_*) \\ &\coloneqq N \int_{\mathcal{X}^N} F(\mu_{\boldsymbol{x}}^N) m_t^N(\mathrm{d}\boldsymbol{x}) + H(m_t^N) - NF(m_*) - NH(m_*), \end{aligned}$$

et examinons son évolution dans le temps. On remarque que dans l'expression ci-dessus, on a utilisé le même symbole $H(\cdot)$ pour la fonctionnelle d'entropie définie pour les mesures de probabilité sur $\mathcal{X}^N$ et $\mathcal{X}$. Comme nous l'avons indiqué précédemment, la dynamique à N particules est en fait linéaire, et l'on a

$$\mathcal{F}^N(m_t^N) = N \int_{\mathcal{X}^N} F(\mu_{\boldsymbol{x}}^N) m_t^N(\mathrm{d}\boldsymbol{x}) + H(m_t^N) = H(m_t^N | m_*^N) + \text{constante},$$

où m_*^N est la mesure invariante des N particules avec une densité

$$m_*^N(\boldsymbol{x}) \propto \exp\big(-NF(\mu_{\boldsymbol{x}}^N)\big).$$

Ainsi, par des calculs classiques, on obtient

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}^N(m_t^N | m_*) = - \sum_{i \in [N]} \int_{\mathcal{X}^N} \left| \nabla_i \log \frac{m_t^N(\boldsymbol{x})}{m_*^N(\boldsymbol{x})} \right|^2 m_t^N(\mathrm{d}\boldsymbol{x}) \eqqcolon -I(m_t^N | m_*^N),$$

où $I(\cdot|\cdot)$ désigne l'information de Fisher relative. L'approche habituelle consiste alors à trouver des conditions pour une inégalité de Sobolev logarithmique uniforme en N pour m_*^N, de sorte que l'on puisse conclure comme suit :

$$I(m_t^N | m_*^N) \gtrsim H(m_t^N | m_*^N) \geqslant \mathcal{F}^N(m_t^N | m_*) - O(1).$$

C'est en effet la méthode de Malrieu [159] et de Carrillo, McCann et Villani [39], et ainsi que des plus récents travaux de Guillin, W. Liu, L. Wu et C. Zhang [100, 99]. Toutefois, les conditions pour l'ISL qui y sont énoncées semblent davantage liées à la convexité de déplacement et ne semblent pas compatibles avec la convexité plate, qui est notre condition structurelle pour les comportements en temps long. Notre innovation consiste à considérer l'information de Fisher conjointe à N particules $I(m_t^N | m_*^N)$ comme la valeur moyenne des informations de Fisher entre les mesures conditionnelles d'*une seule particule*. Cela est possible grâce à la décomposition par composant :

$$\sum_{i\in[N]}\int_{\mathcal{X}^N}\left|\nabla_i \log\frac{m_t^N(\boldsymbol{x})}{m_*^N(\boldsymbol{x})}\right|^2 m_t^N(\mathrm{d}\boldsymbol{x})$$
$$=\sum_{i\in[N]}\int_{\mathcal{X}^{N-1}}\int_{\mathcal{X}}|\nabla_i \log m_t^{N,i|-i}(x^i|\boldsymbol{x}^{-i})+D_mF(\mu_{\boldsymbol{x}}^N,x^i)|^2$$
$$m_t^{N,i|-i}(\mathrm{d}x^i|\boldsymbol{x}^{-i})m_t^{-i}(\mathrm{d}\boldsymbol{x}^{-i}). \quad (15)$$

Ici, $-i$ représente l'ensemble de tous les indices sauf i, c'est-à-dire $[N]\backslash\{i\}$, et $m_t^{N,-i}$, $m_t^{N,i|-i}$ sont respectivement les mesures marginale et conditionnelle définies par

$$m_t^{N,-i}(\boldsymbol{x}^{-i}) \coloneqq \int_{\mathcal{X}} m_t^N(\boldsymbol{x})\,\mathrm{d}x^i,$$
$$m_t^{N,i|-i}(x^i|\boldsymbol{x}^{-i}) \coloneqq \frac{m_t^N(\boldsymbol{x})}{m^{N,-i}(\boldsymbol{x}^{-i})}.$$

En supposant que les mesures de probabilité $\hat{m}$ de la forme suivante

$$\hat{m}(x) \coloneqq \frac{\exp\left(-\frac{\delta F}{\delta m}(m,x)\right)}{\int_{\mathcal{X}}\exp\left(-\frac{\delta F}{\delta m}(m,y)\right)\mathrm{d}y}, \qquad \text{où } m\in\mathcal{P}(\mathcal{X})$$

satisfont une ISL uniforme, on peut (après quelques manipulations sur les mesures) appliquer l'*ISL pour la mesure d'une particule* à l'information de Fisher conditionnelle dans la décomposition (15). On note que cette approche par composant n'est pas entièrement nouvelle, comme elle a déjà été utilisée pour prouver la stabilité d'ISL par tensorisation (voir par exemple [148, Section 5.2]) et l'idée de décomposition est en fait à la base de nombreuses inégalités de concentration adimensionnelles (voir par exemple les discussions sur l'inégalité d'Efron-Stein dans [27, Section 3.1]). L'innovation ici est que la mesure de base m_*^N n'est pas nécessairement tensorisée et on parvient à contrôler les erreurs provenant de la dépendance entre les particules. Enfin, en utilisant la convexité plate de F dans les étapes intermédiaires, on obtient une ISL à N particules avec un terme d'erreur :

$$I(m_t^N|m_*^N) \gtrsim \mathcal{F}^N(m_t^N|m_*) - O(1) \quad (16)$$

et cela permet de conclure par le lemme de Grönwall

$$\mathcal{F}^N(m_t^N|m_*) \leqslant Ce^{-ct}\mathcal{F}^N(m_t^N|m_*) + O(1) = O(Ne^{-ct}+1).$$

On note qu'en remplaçant m_t^N par la mesure invariante m_*^N, on obtient

$$\mathcal{F}^N(m_*^N|m_*) = O(1)$$

et donc (16) implique

$$I(m_t^N|m_*^N) \gtrsim \mathcal{F}^N(m_t^N|m_*) - \mathcal{F}^N(m_*^N|m_*) - O(1) = H(m_t^N|m_*^N) - O(1),$$

ce qui est une *inégalité de log-Sobolev non tendue* pour m_*^N avec des constantes uniformes en N. Pour cette raison, par abus de langage, on dit que l'inégalité (16) est une version non tendue (ou bruitée, discrétisée, défecteuse) de l'ISL non linéaire de champ moyen :

$$I(m|\hat{m}) \gtrsim H(m|\hat{m}) \geqslant \mathcal{F}(m) - \mathcal{F}(m_*) \geqslant H(m|m_*). \tag{17}$$

Les calculs ci-dessus seront expliqués en détail et rigueur dans la démonstration du théorème 1.12 dans le chapitre 1. Le contrôle sur l'énergie libre relative $\mathcal{F}^N(m_t^N|m_*)$ implique ensuite, en raison de la convexité de F et de l'inégalité de Talagrand,

$$W_2^2(m_t^N, m_*^{\otimes N}) \lesssim H(m_t^N|m_*^{\otimes N}) = O(Ne^{-ct} + 1).$$

On combine ceci avec la convergence exponentielle du flot de champ moyen

$$W_2^2(m_t, m_*) \lesssim H(m_t|m_*) = O(e^{-ct})$$

et la borne standard de propagation du chaos qui croît exponentiellement :

$$\frac{1}{N} W_2^2(m_t^N, m_t^{\otimes N}) = O(e^{ct}N^{-\alpha}), \qquad \text{pour un certain } \alpha > 0.$$

Finalement, à travers un argument de triangle, on obtient la borne uniforme souhaitée

$$\sup_{t \geqslant 0} \frac{1}{N} W_2^2(m_t^N, m_t^{\otimes N}) = O(N^{-\alpha'}), \qquad \text{pour un certain } \alpha' > 0.$$

Nous menons également une étude assez détaillée de l'hypercontractivité (inverse) de l'évolution non linéaire (10) et obtenons une *borne inférieure* assez forte sur la densité de m_t à long terme. Cela nous permet de reproduire l'argument du triangle ci-dessus et d'obtenir la borne uniforme sur l'entropie relative :

$$\sup_{t \geqslant 0} \frac{1}{N} H(m_t^N|m_t^{\otimes N}) = O(N^{-\alpha''}), \qquad \text{pour un certain } \alpha'' > 0.$$

Dans le chapitre 2, nous nous tournons vers l'étude du comportement en temps long du flot non linéaire hypoelliptique (14) et de son système de particules associé. À noter que dans ce cas, contrairement à la Langevin de champ moyen suramortie, la convergence exponentielle du flot de champ moyen ne semble même pas avoir été établie, ce qui constitue également la première contribution de ce chapitre. Suivant les intuitions physiques, pour une mesure de probabilité $m \in \mathcal{P}(\mathcal{X} \times \mathbb{R}^d)$, la fonctionnelle d'énergie libre naturelle est la somme des énergies potentielle et cinétique, ainsi que de l'entropie :

$$\mathcal{F}(m) = F(m^X) + \frac{1}{2} \int_{\mathcal{X} \times \mathbb{R}^d} |v|^2 m(\mathrm{d}x\,\mathrm{d}v) + H(m),$$

où m^X est la première loi marginale de m. La particularité de l'hypoellipticité réside dans le fait que, lorsque l'on prend la dérivée temporelle de la fonctionnelle d'énergie libre, on n'obtient pas une information de Fisher complète, mais seulement partielle dans les directions de la vitesse :

$$\begin{aligned}\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} &= -\int_{\mathcal{X}} |\nabla_v \log m_t(x,v) + v|^2 m_t(\mathrm{d}x\,\mathrm{d}v) \\ &= -\int_{\mathcal{X}} \left|\nabla_v \log \frac{m_t(x,v)}{\hat{m}_t(x,v)}\right|^2 m_t(\mathrm{d}x\,\mathrm{d}v),\end{aligned}$$

où $\hat{m}$ est la mesure de probabilité avec la densité

$$\hat{m}(x,v) \propto \exp\biggl(-\frac{\delta F}{\delta m}(m^X,x) - \frac{1}{2}|v|^2\biggr).$$

Une solution pour traiter cette dégénérescence courante dans les modèles cinétiques est d'introduire une quantité tordue, comme Y. Guo [103] et Talay [217] l'ont démontré, et c'est en effet la solution que Villani a trouvée pour la dynamique linéaire de Fokker-Planck [221, Part I]. Dans le cas linéaire, l'énergie potentielle F est donnée par une fonction potentielle :

$$F(m^X) = \int_{\mathcal{X}} U(x) m^X(\mathrm{d}x), \qquad \text{pour une certaine } U\colon \mathcal{X} \to \mathbb{R},$$

et donc $\hat{m}$ est indépendante de m et identique à la mesure d'équilibre, que l'on note m_*. L'idée de Villani est d'introduire une information de Fisher anisotrope (mais toujours définie positive) :

$$\begin{aligned} I_{a,b,c}(m|m_*) = \int_{\mathcal{X}\times\mathbb{R}^d} \biggl(a\biggl|\nabla_v \log \frac{m(z)}{m_*(z)}\biggr|^2 + 2b\nabla_v \log\frac{m(z)}{m_*(z)} \cdot \nabla_x \log \frac{m(z)}{\hat{m}(z)} \\ + c\biggl|\nabla_x \log\frac{m(z)}{m_*(z)}\biggr|^2\biggr) m_t(\mathrm{d}z), \end{aligned}$$

où $z = (x,v)$. En choisissant les bons a, b, c et en effectuant quelques calculs longs, on obtient

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigl(\mathcal{F}(m_t) + I_{a,b,c}(m_t|m_*)\bigr) \lesssim -\int_{\mathcal{X}\times\mathbb{R}^d} \biggl|\nabla_z \log\frac{m_t(z)}{m_*(z)}\biggr|^2 m_t(\mathrm{d}z) =: -I(m_t|m_*).$$

L'information de Fisher pleine du côté droit nous permet de conclure par l'ISL habituelle pour m_* et le phénomène est appelé *hypocoercivité*. Dans notre travail, nous adaptons cette construction de Villani à notre cadre non linéaire, en remplaçant le m_* dans l'information de Fisher anisotrope par $\hat{m}_t$. On doit alors calculer la dérivée temporelle de

$$\mathcal{F}(m_t) + I_{a,b,c}(m_t|\hat{m}_t).$$

La plupart des calculs suivent la ligne des travaux originaux de Villani, puisque le générateur de l'évolution au temps t annihile la mesure $\hat{m}_t$, mais on doit également contrôler le terme supplémentaire provenant de la variation de $\hat{m}_t$. Heureusement, cela ne pose aucun problème pour l'hypocoercivité et nous obtenons la convergence exponentielle du flot de champ moyen (voir également la démonstration du théorème 2.2).

Pour le système de N particules associé, puisque la dynamique est linéaire, on peut appliquer directement le formalisme de Villani, et le point essentiel est d'avoir une hypocoercivité uniforme en N. Plus précisément, on a

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigl(H(m^N_t|m^N_*) + I_{a,b,c}(m^N_t|m^N_*)\bigr) \leqslant -\kappa I(m^N_t|m^N_*)$$

pour certains a, b, c et $\kappa > 0$ qui ne dépendent pas de N. Dans le cas cinétique, on a également une ISL discrétisée similaire à (16), ce qui permet d'obtenir de nouveau

$$\mathcal{F}^N(m^N_t) - N\mathcal{F}(m_*) \leqslant O(Ne^{-ct} + 1).$$

On explique les arguments en détail dans la démonstration du théorème 2.3.

Enfin, nous travaillons également sur les propriétés de régularisation à court terme de la dynamique hypoelliptique. En adaptant la méthode de *couplage par changement de mesure* développée par Guillin, P. Ren et F.-Y. Wang dans une série de travaux [102, 227, 193], nous obtenons des inégalités de log-Harnack sans dimension pour le flot de champ moyen et le système de particules, ce qui conduit à la régularisation de la distance de Wasserstein à l'entropie relative. Ensuite, nous adaptons la fonctionnelle de Hérau [109] à notre cadre pour obtenir la régularisation de l'entropie relative à l'information de Fisher relative. Enfin, nous combinons les convergences exponentielles à long terme, les régularisations à court terme et la propagation standard exponentiellement croissante du chaos pour dériver une borne de propagation du chaos uniforme en temps, sans nécessiter de régularité des données initiales (voir le théorème 2.6).

Aperçu des chapitres 3 et 4

ISL et applications

L'objectif principal des chapitres 3 et 4 est d'obtenir des critères d'ISL pour les mesures de probabilité liées à un processus de diffusion sans avoir directement accès à leurs densités. Dans le chapitre 3, nous souhaitons montrer une ISL pour la mesure *stationnaire* de la diffusion homogène en temps suivante

$$\partial_t m_t = \frac{\sigma^2}{2}\Delta m_t - \nabla \cdot (b m_t),$$

et dans le chapitre 4, nous permettons à la dérive b de dépendre du temps et souhaitons montrer une ISL *uniforme en temps* pour m_t résolvant l'inhomogène

$$\partial_t m_t = \frac{\sigma^2}{2}\Delta m_t - \nabla \cdot (b_t m_t). \tag{18}$$

Notons d'abord que le premier problème est en fait plus ou moins inclus dans le second, car si l'on peut montrer une ISL uniforme en temps pour m_t, $t \geqslant 0$, et sait que m_t converge faiblement vers une mesure stationnaire m_*, alors m_* satisfait également une ISL. Nous travaillerons donc principalement dans le cadre parabolique dans le reste de l'aperçu. Deuxièmement, si la dérive dans le premier problème est un gradient :

$$b(x) = -\nabla U(x),$$

alors la diffusion est symétrique et l'on sait que la mesure stationnaire a une densité proportionnelle à $\exp\bigl(-2U(x)/\sigma^2\bigr)$. Dans ce cas, par le critère classique de Bakry-Émery et les résultats de perturbation, on sait déjà comment montrer une ISL pour une grande classe de U. Ainsi, l'intérêt du chapitre 3 réside dans le cas non symétrique. Enfin, dans le cas où b satisfait une forte monotonie :

$$\bigl(b_t(x) - b_t(y)\bigr) \cdot (x - y) \leqslant -\kappa |x - y|^2, \qquad \text{pour un certain } \kappa > 0,$$

en répétant l'argument classique de Malrieu [159] (voir aussi le début du chapitre 4), on peut propager l'ISL uniformément dans le temps, et le second problème est résolu. Nous nous concentrerons donc sur les cas où cette forte monotonie est absente.

Comme nous l'avons mentionné ci-dessus, nous proposons deux classes de tels critères basés sur deux méthodes différentes. La première méthode est basée sur un travail récent de Monmarché [168], où il a montré que, si b est assez régulier et si b n'est non monotone qu'à l'intérieur d'un ensemble compact, alors il existe σ_0 tel que pour tout $\sigma \geqslant \sigma_0$, le processus de diffusion lié à (18) est contractant en Wasserstein de L^2. Autrement dit, pour deux solutions μ_t, ν_t de (18), on a

$$W_2(\mu_t, \nu_t) \leqslant M e^{-\lambda t} W_2(\mu_0, \nu_0), \qquad \text{pour un certain } \lambda > 0. \tag{19}$$

Monmarché présente deux démonstrations de ce résultat dans son travail cité et la démonstration probabiliste est basée sur le couplage synchrone des processus de diffusion et un coût de transport modifié équivalent à la distance euclidienne au carré. La condition de haute température est crucialement utilisée pour construire un tel coût de transport. Ensuite, par des arguments standards (en fait, une version L^2 du résultat de propagation de Malrieu [159]), la contraction (19) mène à une inégalité de Poincaré uniforme pour m_t. On observe ensuite que, une fois que nous avons un contrôle uniforme en temps sur $x \mapsto x \cdot b_t(x)$ et ∇b_t, on peut obtenir respectivement une borne de moment gaussienne et une inégalité de Harnack (de F.-Y. Wang) qui sont toutes les deux uniformes dans le temps. Ces deux résultats conduisent à une hypercontractivité uniforme en temps, ce qui équivaut à une inégalité de log-Sobolev non tendue uniforme en temps. Ensuite, en combinant l'inégalité de Poincaré et l'ISL défectueuse, on obtient l'ISL uniforme en temps souhaitée.

La seconde méthode repose sur des estimations directes sur la densité de la solution m_t de (18) et peut sembler brutale pour les lecteurs versés dans les inégalités fonctionnelles. Pour simplifier, on fixe $\sigma = \sqrt{2}$ dans cet alinéa. Supposons qu'il existe une mesure de référence μ_0 satisfaisant une ISL qui est aussi stationnaire par rapport à la dérive a_0, c'est-à-dire,

$$\Delta \mu_0 - \nabla \cdot (a_0 \mu_0) = 0.$$

Notons la différence des dérives par $g_t \coloneqq b_t - a_0$ et la densité relative logarithmique par $u_t \coloneqq \log m_t/\mu_0$. Selon les EDP respectives de m_t et μ_0, on trouve que u_t résout l'équation de Hamilton-Jacobi-Bellman (HJB)

$$\partial_t u_t = \Delta u_t + |\nabla u_t|^2 + \tilde{b}_t \cdot \nabla u_t + \varphi_t,$$

où les coefficients sont définis par

$$\begin{aligned} \tilde{b}_t &\coloneqq 2\nabla \log \mu_0 - b_t, \\ \varphi_t &\coloneqq -\nabla \cdot g_t + g_t \cdot \nabla \log \mu_0. \end{aligned}$$

On dit que la dérive $\tilde{b}$ est *faiblement semi-monotone* si

$$\bigl(\tilde{b}_t(x) - \tilde{b}_t(y)\bigr) \cdot (x-y) \leqslant -\kappa(|x-y|)|x-y|^2,$$

pour une certaine $\kappa\colon (0,\infty) \to \mathbb{R}$ telle que $\liminf_{r\to\infty} \kappa(r) > 0$ et $r \mapsto r|\kappa(r)|$ est intégrable près de 0. Dans un travail récent [61], Conforti a démontré que si $\tilde{b}$ est faiblement semi-monotone et φ est lipschitzienne, alors il existe une borne uniforme en temps sur le gradient de u_t. Ainsi, selon le résultat de perturbation log-lipschitzienne d'Aida et Shigekawa [1], le flot de mesures $m_t \propto \mu_0 \exp(u_t)$ satisfait une ISL uniforme en temps. La méthode de Conforti pour cette estimation du gradient est probabiliste : il utilise le *couplage par réflexion* pour des processus de diffusion contrôlés et montre la contraction en distance W_1, ce qui conduit à l'estimation du gradient uniforme en temps. La version non contrôlée du couplage par réflexion semble avoir été développée pour la première fois par Lindvall et Rogers [152] dans les années 1980. Ce couplage a été généralisé aux diffusions sur des variétés par Kendall [130] et ensuite utilisé pour dériver des estimations de gradient pour l'équation de la chaleur par Cranston [65]. La semi-convexité faible, avec le couplage par réflexion, a été exploitée par M.-F. Chen et F.-Y. Wang [47] dans les années 1990 pour estimer le trou spectral du générateur de diffusion et les effets régularisants à court terme sont dérivés par Priola et F.-Y. Wang dans [187]. Porretta et Priola ont ensuite montré l'effet de régularisation pour le flot de HJB non linéaire dans [186] en utilisant le principe de comparaison purement analytique entre les solutions de viscosité. Le travail plus récent d'Eberle [83] a ravivé cette méthode car il a attiré beaucoup d'attention des communautés de statistiques et d'apprentissage automatique. Nous remarquons que le travail de Conforti cité ci-dessus a apporté deux contributions vitales au chapitre 4 : d'abord, on obtient l'estimation *en temps long* dans le cas de HJB ; et ensuite, on démontre également l'estimation du hessien (c'est-à-dire sur $\nabla^2 u_t$). Nous commenterons en particulier sur la deuxième contribution ci-dessous.

Dans le reste du chapitre 4, nous discutons quelques exemples qui vérifient les deux critères présentés ci-dessus et appliquons l'ISL uniforme en temps pour obtenir la propagation du chaos locale uniforme en temps pour la dynamique de McKean-Vlasov avec des potentiels d'interaction non convexes, ce qui n'est pas incluse dans l'article de Lacker et Le Flem [142]. La propagation du chaos locale sera discutée plus en détail dans l'aperçu du chapitre suivant. Cependant, l'application la plus intéressante de notre méthode est peut-être le *modèle de vortex en 2D dans l'espace entier* présenté à la fin du chapitre. Le modèle de vortex en 2D est une formulation probabiliste des équations de Navier-Stokes incompressibles en 2D et nous renvoyons le lecteur à l'article d'exposition [205] pour plus de détails. Dans ce modèle, le flot de champ moyen (1) suit la dérive de McKean-Vlasov

$$b(m,x) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{(x-y)^\perp}{|x-y|^2} m(\mathrm{d}y),$$

où le symbole $\perp$ représente la rotation en 2D : $(x_1,x_2)^\perp = (-x_2,x_1)$. En d'autres termes, la dérive vérifie $b(m,x) = (K \star m)(x)$ où K est le noyau de Biot-Savart. Récemment, Jabin et Z. Wang [124] ont montré la propagation du chaos globale en temps pour ce modèle et Guillin, Le Bris et Monmarché [98] l'ont améliorée en une borne uniforme de la propagation du chaos. Cependant, puisque la méthode de Jabin et Z. Wang est basée sur un *principe d'unicité faible-fort*, elle nécessite une régularité assez forte du flot de champ moyen. Pour être plus précis, il faut contrôler la norme L^∞ de $\nabla \log m_t$ et $\nabla^2 \log m_t$. Ceci est assez difficile à établir dans l'espace entier car on ne peut pas avoir de borne inférieure globale sur la densité m_t. Pour cette raison, les deux travaux cités ci-dessus ne traitent que le modèle de vortex en 2D périodique sur le tore bidimensionnel. Dans le chapitre 4, on montre qu'en ajoutant un confinement quadratique supplémentaire, c'est-à-dire, en laissant

$$b(m,x) = -\kappa x + (K \star m)(x),$$

on peut utiliser la méthode de HJB pour obtenir des bornes L^∞ sur

$$\nabla \log \frac{m_t}{m_*},\ \nabla^2 \log \frac{m_t}{m_*}, \qquad \text{où } m_* \propto \exp\biggl(-\frac{\kappa|x|^2}{2}\biggr).$$

De plus, ces bornes convergent vers zéro de façon exponentielle et cela permet de montrer la propriété de *génération du chaos* pour le modèle de vortex en 2D. Les démonstrations de telles bornes L^∞ sont modérément longues. En raison de la singularité du noyau de Biot-Savart K, on s'appuie sur une procédure de bootstrap parabolique pour gagner progressivement en régularité sur les coefficients $\tilde{b}$, φ, et l'on a besoin à la fois de la contraction à long terme de Conforti et de la régularisation à court terme de Porretta, Priola et F.-Y. Wang. Pour conclure la discussion sur les vortex en 2D, on note que, après que le chapitre 4 était apparu en prépublication [170], Rosenzweig et Serfaty ont mis en ligne leur prépublication [201] où ils montrent que les modèles de vortex en 2D avec et sans confinement quadratique sont équivalents par une transformation d'échelle. Ainsi, notre méthode peut également s'appliquer au modèle sur l'espace entier sans confinement.

Enfin, nous mentionnons que dans le chapitre 3, on développe également un critère d'ISL pour la mesure stationnaire d'une diffusion cinétique. Il est basé sur la méthode de HJB et l'étape la plus importante est de construire une contraction de Wasserstein pour les processus de diffusion cinétiques contrôlés. L'énoncé et la démonstration sont présentés à la fin du chapitre. La méthode repose sur un couplage mixte, comprenant à la fois des parties synchrone et réfléchie, et un coût de transport tordu (habituel pour les modèles cinétiques) motivé par la construction d'Eberle, Guillin et Zimmer [84]. On peut considérer ce résultat comme une généralisation (ou même une amélioration à certains égards) des travaux récents de Kazeykina, Z. Ren, X. Tan et J. Yang [128] et de Schuh [206].

Aperçu du chapitre 5

Taille du chaos pour les dynamiques singulières

Dans le chapitre 5, nous étudions une propriété fine du grand système de particules de champ moyen appelée *propagation du chaos locale*. Bien que nous ayons étudié la propagation du chaos quantitative dans les chapitres précédents, les résultats obtenus concernent uniquement les distances entre le système de particules et le système de champ moyen tensorisé dans son ensemble, par exemple, la distance de Wasserstein $W_2^2(m_t^N, m_t^{\otimes N})$ ou l'entropie relative $H(m_t^N|m_t^{\otimes N})$. Dans ce chapitre, au lieu d'étudier ces distances *globales*, on observe uniquement les k premières particules du système des N particules (3) et les compare avec le flot de champ moyen tensorisé k fois. Pour justifier le fait de ne considérer que les k premières particules et non d'autres ensembles de k particules, nous devons bien entendu supposer l'échangeabilité dans le système de N particules, et cette hypothèse sera en vigueur tout au long du chapitre. On suppose également que l'interaction de champ moyen dans la dérive prend la forme de noyau suivante :

$$b(m, x) = \int_{\mathcal{X}} K(x, y) m(\mathrm{d}y).$$

Rappelons que la loi du sous-système de k particules au temps t est notée $m_t^{N,k}$, ou en d'autres termes,

$$m_t^{N,k} := \mathrm{Loi}(X_t^1, \ldots, X_t^k),$$

où la dépendance de N dans le membre de droite est implicite. La question soulevée consiste donc à trouver une borne précise entre les deux mesures de probabilité $m_t^{N,k}$ et $m_t^{\otimes k}$. C'est une version quantitative de la condition de chaos (6).

Pour cette raison, la question est également appelée *taille du chaos* dans les littératures récentes. Voir les travaux de Paul, Pulvirenti et Simonella [182], de

Duerinckx [78], et de Bernou et Duerinckx [20]. Un point commun des approches de ces travaux est que les auteurs décomposent la loi des N particules en une somme combinatoire de fonctions de corrélation connexes (ou cumulants), la décomposition étant appelée développement en clusters, et ils étudient l'évolution des fonctions de corrélation le long de la dynamique. En fonction de l'interaction spécifique de champ moyen, les corrélations entre les particules peuvent être engendrées par des collisions ou à travers la dérive et l'aléa peut provenir de l'initialisation et aussi du bruit dynamique. Après avoir estimé la taille des cumulants, ils reviennent alors au problème de la taille du chaos et obtiennent

$$\|m_t^{N,k} - m_t^{\otimes k}\| = O\left(\frac{k^2}{N}\right),$$

où $\|\cdot\|$ désigne une norme fonctionnelle appropriée. Grossièrement parlant, le facteur k^2 provient du comptage du nombre de paires parmi les k premières particules. Ce facteur ne peut pas être réduit dans cette approche, à moins qu'une certaine annulation ne se produise, ce qui peut résulter de l'orthogonalité.

Une approche complètement différente est développée dans le travail récent de Lacker [140], où il considère directement l'évolution des erreurs entre $m_t^{N,k}$ et $m_t^{\otimes k}$, mesurées en termes d'entropie relative

$$H_t^k \coloneqq H(m_t^{N,k}|m_t^{\otimes k}).$$

La dynamique de la mesure $m_t^{N,k}$ est décrite par la hiérarchie de BBGKY et implique la marginale d'ordre supérieur $m_t^{N,k+1}$, et donc l'évolution de H_t^k devrait également impliquer des quantités d'ordre supérieur. En effet, dans la dernière étape de la démonstration de Lacker, l'équation dynamique s'écrit

$$\frac{\mathrm{d}H_t^k}{\mathrm{d}t} \leqslant M\frac{k^2}{N^2} + M(H_t^{k+1} - H_t^k),$$

où M est une constante liée à la force du noyau d'interaction de champ moyen K. En résolvant le système d'inégalités ci-dessus, on obtient la borne globale en temps $H_t^k = O(k^2/N^2)$, et en termes de distance de norme,

$$\|m_t^{N,k} - m_t^{\otimes k}\|_{\mathrm{TV}} = O\left(\frac{k}{N}\right)$$

ce qui améliore les résultats ci-dessus par un facteur de k. Cette borne est optimale car elle peut être atteinte par un simple exemple gaussien. Par la suite, cette méthode a été étendue au cas uniforme en temps dans le régime d'interaction faible par Lacker et Le Flem [142] et la borne optimale pour le chaos d'ordre supérieur est obtenue par Hess-Childs et Rowan [111]. Nous remarquons que la méthode de Lacker repose crucialement sur le bruit brownien pour contrôler la croissance de H_t^k et c'est peut-être la raison du gain de facteur k par rapport aux approches plus combinatoires ci-dessus.

Une limitation commune des travaux précédents sur la taille du chaos est que l'on nécessite une forte hypothèse de régularité (au moins L^∞) sur le noyau d'interaction K, excluant ainsi le modèle intéressant de vortex en 2D où le noyau est de Biot-Savart :

$$K(x,y) = \frac{(x-y)^\perp}{2\pi|x-y|^2}.$$

Le but du chapitre 5 est précisément de surmonter cette limitation et de démontrer la borne optimale de la taille du chaos pour le système de particules de vortex en 2D, c'est-à-dire, $H_t^k = O(k^2/N^2)$. En combinant les techniques de Jabin-Z. Wang et de Lacker, on montre que l'évolution de l'entropie relative vérifie

$$\frac{\mathrm{d}H_t^k}{\mathrm{d}t} \leqslant -c_1 I_t^k + c_2 I_t^{k+1} + M\biggl(H_t^k + \frac{k^2}{N^2}\biggr) + M(H_t^{k+1} - H_t^k),$$

où I_t^k, I_t^{k+1} sont des informations de Fisher relatives. En particulier, on a

$$I_t^k \coloneqq \sum_{i\in[k]} \int_{\mathcal{X}^k} \biggl|\nabla_i \log \frac{m_t^{N,k}(\boldsymbol{x}^{[k]})}{m_t^{\otimes k}(\boldsymbol{x}^{[k]})}\biggr|^2 m_t^{N,k}(\mathrm{d}\boldsymbol{x}^{[k]}).$$

La principale différence par rapport au travail de Lacker est bien sûr l'information de Fisher *positive d'ordre supérieur* I_t^{k+1}, qui provient de la singularité du noyau K. Résoudre le système d'inégalités différentielles dans le cas $c_2 < c_1$ est la principale innovation technique du chapitre. Remarquons que la condition $c_2 < c_1$ correspond au fait que la norme $W^{-1,\infty}$ du noyau K est inférieure à 1, donc notre résultat est valide pour des interactions de vortex faibles, ou de manière équivalente, pour des vortex dans un régime de haute température. L'idée principale de la démonstration est de considérer un mélange pondéré des entropies d'ordre $\geqslant k$:

$$Z_t^k \coloneqq \sum_{i=k}^N a_{k,i} H_t^i, \qquad \text{où } a_{k,i} \geqslant 0 \text{ et } a_{k,k} = 1.$$

En choisissant les coefficients appropriés $a_{k,i}$, on peut annuler toutes les informations de Fisher dans la dynamique de Z_t^k et retrouver le système original de Lacker. On en déduit ainsi $Z_t^k = O(k^2/N^2)$ et on peut conclure par $H_t^k \leqslant Z_t^k$. En utilisant les idées de [98], on améliore également la borne globale en temps sur la taille du chaos pour obtenir une borne uniforme. On en discute également certaines conséquences. Par exemple, en tirant parti de l'injection de L^d dans $W^{-1,\infty}$ [28], on peut démontrer une borne optimale et globale en temps pour la taille du chaos pour les interactions en L^d de toute intensité. On utilise également une approche en L^2 (plutôt qu'en entropie) pour la taille du chaos dans le cas des interactions de vortex afin de lever la restriction sur la force d'interaction, mais malheureusement on n'obtient un résultat que pour un temps fini.

Aperçu des chapitres 6 et 7

Jeu fictif et auto-interaction

Dans les chapitres 6 et 7, nous étudions des dynamiques de champ moyen alternatives qui approchent le minimiseur du problème d'optimisation de champ moyen régularisé par l'entropie (11) en temps long. La dynamique d'intérêt dans le chapitre 6 est le jeu fictif entropique défini de la manière suivante :

$$\partial_t m_t = \alpha(\hat{m}_t - m_t), \qquad \text{où } \hat{m}_t \propto \exp\biggl(-\frac{\delta F}{\delta m}(m,\cdot)\biggr). \tag{20}$$

La définition des dynamiques ci-dessus est motivée par l'algorithme de jeu fictif, d'abord proposé par Brown [34] dans le cadre d'un jeu à deux personnes. Dans un jeu symétrique à deux personnes avec un espace d'états continu, on note les états des deux joueurs par x, y respectivement, et la condition d'équilibre de Nash s'écrit

$$x_* \in \mathrm{MR}(y_*), \qquad y_* \in \mathrm{MR}(x_*),$$

où $\mathrm{MR}(\cdot)$ est l'ensemble des meilleures réponses donné l'état de l'adversaire. Brown propose que les deux joueurs suivent les dynamiques discrètes respectives

$$\begin{aligned} x_{t+1} &= \frac{t}{t+1}x_t + \frac{1}{t+1}a_t, \qquad \text{où } a_t \in \mathrm{MR}(y_t),\\ y_{t+1} &= \frac{t}{t+1}y_t + \frac{1}{t+1}b_t, \qquad \text{où } b_t \in \mathrm{MR}(x_t), \end{aligned}$$

et s'attend à ce que (x_t, y_t) converge vers un certain équilibre de Nash (x_*, y_*) à long terme. Pour comprendre les intuitions derrière nos dynamiques de jeu fictif entropique, remarquons que par le calcul variationnel, la condition du premier ordre du problème d'optimisation (11) est

$$\frac{\delta F}{\delta m}(m_*, x) + \log m_*(x) = \text{constante}.$$

Selon la définition de la mesure $\hat{m}$, la condition ci-dessus est équivalente à

$$m_* = \hat{m}_*.$$

C'est une condition d'équilibre de Nash pour le jeu à une personne (ou auto-jeu) si la fonction $m \mapsto \hat{m}$ est interprétée comme la fonction de meilleure réponse. Et si l'on remplace le facteur $1/t$ dans les dynamiques de Brown par une échelle exponentielle, et considérons la version continue, les dynamiques de jeu fictif correspondantes sont exactement (20). Nous remarquons que Cardaliaguet et Hadikhanloo ont aussi utilisé cette idée afin de trouver des solutions aux jeux à champ moyen [36], ce qui peuvent également être formulés comme un problème du point fixe.

En général, on ne peut pas garantir la convergence de l'algorithme de jeu fictif, mais dans le cas des jeux de potentiel, on peut souvent trouver des fonctions de Lyapunov qui diminuent le long des dynamiques. Pour le jeu fictif entropique, on calcule la dérivée temporelle de la fonctionnelle d'énergie libre (qui est la fonctionnelle à optimiser) le long des dynamiques, et l'on trouve

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\alpha\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big).$$

Puisque l'on suppose toujours que la fonction d'énergie F est convexe au sens plat, on peut déjà utiliser l'inégalité de sandwich de l'entropie (1.40) dans le chapitre 1 :

$$H(m_t|\hat{m}_t) \geqslant \mathcal{F}(m_t) - \mathcal{F}(m_*).$$

et la convergence exponentielle suit. Cependant, ce n'est pas l'approche que nous avons adoptée dans le chapitre 6, en partie parce que nous n'étions pas familiers avec une telle inégalité de sandwich lorsque le papier correspondant [49] a été écrit. (Cette inégalité de sandwich a déjà été utilisée dans [56, 178] pour démontrer la convergence exponentielle de la dynamique de Langevin de champ moyen suramortie à cette époque.) Au lieu de cela, on prend à nouveau la dérivée temporelle de

$H(m_t|\hat{m}_t)$ et trouve que, au moins formellement,

$$\frac{\mathrm{d}H(m_t|\hat{m}_t)}{\mathrm{d}t} = -\alpha\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big) \\ - \alpha \iint_{\mathcal{X}^2} \frac{\delta^2 F}{\delta m^2}(m_t, x, y)(\hat{m}_t - m_t)^{\otimes 2}(\mathrm{d}x\,\mathrm{d}y).$$

Le dernier terme est négatif en raison de la convexité de F. On obtient donc

$$\frac{\mathrm{d}H(m_t|\hat{m}_t)}{\mathrm{d}t} \leqslant -\alpha H(m_t|\hat{m}_t), \\ \frac{\mathrm{d}H(m_t|\hat{m}_t)}{\mathrm{d}t} \leqslant \frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t}.$$

Et par quelques calculs élémentaires, on conclut que $\mathcal{F}(m_t)$ diminue exponentiellement. Le calcul du second ordre ci-dessus est intéressant en soi car il ressemble à la démonstration d'Otto-Villani [180] du critère de Bakry-Émery, que nous esquissons ici de manière minimaliste. Supposons que l'on a une dynamique de Langevin suramortie, engendrée par $\Delta - \nabla U \cdot \nabla$ pour une certaine $U \colon \mathbb{R}^d \to \mathbb{R}$ satisfaisant $\nabla^2 U \succeq \rho$ avec $\rho > 0$. Notons par m_* la mesure invariante proportionnelle à $\exp(-U)$ et m_t le flot de mesure associé. Posons également pour simplifier $H_t \coloneqq H(m_t|m_*)$ et $I_t \coloneqq I(m_t|m_*)$. Otto et Villani ont calculé que

$$\frac{\mathrm{d}H_t}{\mathrm{d}t} = -I_t, \\ \frac{\mathrm{d}I_t}{\mathrm{d}t} \leqslant -2\rho I_t.$$

Puisque l'on sait que $\lim_{t\to\infty} H_t = 0$, on a

$$H_0 = \int_0^\infty I_t\,\mathrm{d}t \leqslant \int_0^\infty I_0 e^{-2\rho t}\,\mathrm{d}t = \frac{I_0}{2\rho}.$$

Comme la valeur initiale du flot est arbitraire, on a établi l'inégalité de log-Sobolev, ce qui conduit à la convergence exponentielle de l'entropie relative. Ainsi, dans le jeu fictif entropique, l'énergie libre $\mathcal{F}$ joue le rôle de l'entropie dans Otto-Villani, et $H(m_t|\hat{m}_t)$ joue le rôle d'information de Fisher.

Malgré la simplicité du jeu fictif entropique, une difficulté numérique importante n'est pas prise en compte dans l'analyse ci-dessus. À chaque étape t, on doit calculer la meilleure réponse à m_t, à savoir $\hat{m}_t \propto \exp\big(-\frac{\delta F}{\delta m}(m_t, \cdot)\big)$, et cela se fait généralement par des méthodes de Monte-Carlo : par exemple, on lance des particules à partir d'une distribution initiale et les laissons évoluer selon la dynamique de Langevin suramortie. Après un temps suffisamment long, avec un nombre suffisamment grand de particules, on peut échantillonner la mesure $\hat{m}_t$ avec une précision arbitraire. On appelle cette étape itération intérieure dans le chapitre 6. Cependant, nous n'y abordons pas la complexité algorithmique de cette itération.

C'est la raison pour laquelle, dans le chapitre 7, nous nous tournons vers la dynamique suivante :

$$\mathrm{d}X_t = -D_m F(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \\ \mathrm{d}m_t = \lambda(t)(\delta_{X_t} - m_t)\,\mathrm{d}t. \tag{21}$$

Ici, $\lambda\colon [0,\infty) \to (0,\infty)$ est à déterminer et m_t n'est plus la loi de la particule X_t, mais une mesure d'occupation pondérée de la particule selon la deuxième équation :

$$m_t = e^{-\int_0^t \lambda(s)\,\mathrm{d}s} m_0 + \int_0^t \lambda(s) e^{-\int_s^t \lambda(u)\,\mathrm{d}u} \delta_{X_s}\,\mathrm{d}s.$$

Le terme de dérive de la particule au temps t dépend donc de son histoire sur l'intervalle $[0,t]$ et pour cette raison, la dynamique est dite *auto-interagissante*. Ce type de dynamique a déjà été étudié par Cranston et Le Jan [66], Raimond [188], et Benaïm, Ledoux et Raimond [15]. L'article récent de Du, Jiang et J. Li [75] aborde l'utilité de telles dynamiques dans l'échantillonnage. Pour l'instant, on fixe $\lambda(t) = \lambda > 0$. Remarquons que chacun des deux composants dans (21) a une échelle de temps naturelle. Si l'argument de mesure m_t est figé, le premier composant suit une Langevin suramortie linéaire et l'échelle de temps est le temps de mélange pour un tel processus. De même, en fixant l'argument X_t dans la deuxième équation, on trouve que l'échelle de temps du deuxième composant est $1/\lambda$. Sous la limite $\lambda \to 0$, la deuxième échelle de temps devient beaucoup plus grande que la première, si bien que la distribution du premier argument se stabilise rapidement vers l'état stationnaire $\hat{m}_t$ avant que le second argument ne change de manière significative. Et puisque, par le théorème de Birkhoff, la masse de Dirac δ_{X_t} moyennée sur un intervalle suffisamment long est proche de l'état stationnaire $\hat{m}_t$, on espère que sur le long terme, la dynamique auto-interagissante devrait être décrit effectivement par le jeu fictif entropique :

$$\mathrm{d}m_t = \lambda(\hat{m}_t - m_t)\,\mathrm{d}t,$$

qui converge vers m_* lorsque $t \to \infty$. D'autre part, sous la limite $\lambda \to \infty$, le second argument m_t devient très proche de la masse de Dirac δ_{X_t}, donc la dynamique devrait se rapprocher de la dynamique linéaire

$$\mathrm{d}X_t = -D_m F(\delta_{X_t}, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t.$$

Ce processus de Markov se stabilise rapidement, mais sa mesure d'équilibre, étant proportionnelle à $\exp\bigl(-F(\delta_x)\bigr)\,\mathrm{d}x$, est a priori différente de notre objectif m_*. On se retrouve donc dans une situation similaire au compromis biais-variance classique une fois que l'on fait le lien entre le taux de relaxation et l'inverse de la variance.

Nous étudions quantitativement ce compromis dans le chapitre 7. Nous prenons un $\lambda > 0$ fixe et étudions d'abord le taux de convergence du processus auto-interagissant (21). Remarquons que dans ce cas, le processus est de Markov homogène dans un espace d'état infini-dimensionnel avec un bruit hautement dégénéré, donc prouver son ergodicité est généralement une tâche non triviale. Cependant, grâce à la forte contractivité dans le second argument, nous sommes en mesure d'établir une contractivité exponentielle explicite en distance de Wasserstein par un couplage mixte similaire au couplage cinétique d'Eberle, Guillin et Zimmer [84]. Le taux de contraction obtenu se détériore lorsque λ diminue, ce qui est à prévoir. Notamment, la condition structurelle que F est convexe au sens plat n'est pas du tout utilisée pour la relaxation du processus de Markov. En conséquence, on sait que la mesure stationnaire du processus de Markov (21) existe et est unique, ce que nous notons par P^λ.

Nous étudions ensuite le biais entre la mesure stationnaire P^λ et la cible $m_* \otimes \delta_{m_*}$, où, rappelons-le, m_* est la mesure invariante pour la dynamique de Langevin de champ moyen (9) ou la solution du problème d'optimisation de champ moyen (11). Pour procéder, on suppose que la dépendance de champ moyen est cylindrique :

$$F(m) = \Phi(\langle \ell, m\rangle) = \Phi\biggl(\int_{\mathcal{X}} \ell(x) m(\mathrm{d}x)\biggr)$$

pour un $\ell\colon \mathcal{X} \to \mathbb{R}^D$ et une fonction convexe $\Phi\colon \mathbb{R}^D \to \mathbb{R}$. Ici, la convexité de Φ implique la convexité au sens plat de F en tant que fonctionnelle de champ moyen. Le processus auto-interagissant (21) peut alors être réduit au système projeté

$$\begin{aligned} \mathrm{d}X_t &= -\nabla_x V(Y_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \\ \mathrm{d}Y_t &= \lambda\bigl(\ell(X_t) - Y_t\bigr)\,\mathrm{d}t, \end{aligned} \tag{22}$$

où les variables se correspondent de manière suivante :

$$\begin{aligned} Y_t &= \langle \ell, m_t\rangle, \\ V(y,x) &= \nabla\Phi(y)\cdot\ell(x). \end{aligned}$$

Notons $\rho \coloneqq \rho^\lambda$ l'image de la mesure P^λ par l'application

$$(x, m) \mapsto (x, \langle \ell, m\rangle).$$

Par construction, la mesure ρ est invariante par rapport à la dynamique réduite (22), et résout l'équation stationnaire :

$$\Delta_x\rho + \nabla_x\cdot\bigl(\nabla_x V(y,x)\rho\bigr) - \lambda\nabla_y\cdot\Bigl(\bigl(\ell(x) - y\bigr)\rho\Bigr) = 0.$$

En utilisant l'équation ci-dessus et une inégalité de log-Sobolev uniforme, on démontre l'estimation de L^1 sur l'entropie conditionnelle suivante :

$$\int_{\mathbb{R}^D} H\bigl(\rho^{1|2}(\cdot|y)\big|\hat{m}_y\bigr)\rho^2(\mathrm{d}y) = O(\lambda), \tag{23}$$

où $\rho^{1|2}$ et ρ^2 sont respectivement les mesures conditionnelle et marginale définies formellement par

$$\begin{aligned} \rho^2(y) &\coloneqq \int_{\mathcal{X}} \rho(x,y)\,\mathrm{d}x, \\ \rho^{1|2}(x|y) &\coloneqq \frac{\rho(x,y)}{\rho^2(y)}, \end{aligned}$$

et $\hat{m}_y$ est la mesure de Gibbs qui a pour densité

$$\hat{m}_y(x) \propto \exp\bigl(-V(y,x)\bigr).$$

L'estimation (23) indique qu'en moyenne, $\rho^{1|2}(\cdot|y)$ est proche de $\hat{m}_y$. Désignons la projection cylindrique de la mesure cible par $y_* \coloneqq \langle \ell, m_*\rangle$. On remarque que

$$\begin{aligned} &\int_{\mathbb{R}^D} \bigl(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\bigr)\rho^2(\mathrm{d}y) \\ &\quad= -\int_{\mathcal{X}\times\mathbb{R}^d} \bigl(V(y,x) - V(y_*,x)\bigr)(\hat{m}_y - m_*)(\mathrm{d}x)\rho^2(\mathrm{d}y) \\ &\quad= -\int_{\mathcal{X}\times\mathbb{R}^d} \bigl(V(y,x) - V(y_*,x)\bigr)\bigl(\rho^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\bigr)\rho^2(\mathrm{d}y) + O(\sqrt{\lambda}), \end{aligned}$$

où pour la dernière égalité, on effectue le changement de mesure $\hat{m}_y \to \rho^{1|2}(\cdot|y)$ et contrôle l'erreur par l'estimation de l'entropie (23) et une inégalité de transport (Talagrand, Pinsker ou Bolley-Villani [25] selon l'hypothèse sur V). En utilisant la forme du potentiel $V(y,x) = \nabla\Phi(y)\cdot\ell(x)$ et la convexité de Φ, on peut démontrer que

$$\int_{\mathcal{X}\times\mathbb{R}^d} \big(V(y,x) - V(y_*,x)\big)\big(\rho^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y) \geqslant 0.$$

On obtient donc

$$\int_{\mathbb{R}^D} \big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y) = O(\sqrt{\lambda}).$$

En utilisant à nouveau l'inégalité de Talagrand, on trouve

$$\int_{\mathbb{R}^D} W_1\big(\rho^{1|2}(\cdot|y), m_*\big)\rho^2(\mathrm{d}y) \\ \leqslant \int_{\mathbb{R}^D}\Big(W_1\big(\rho^{1|2}(\cdot|y),\hat{m}_y\big) + W_1(\hat{m}_y, m_*)\Big)\rho^2(\mathrm{d}y) = O(\lambda^{1/4}).$$

Cela indique déjà que les mesures P^λ et $m_*\otimes\delta_{m_*}$, projetées dans la direction X, sont proches l'une de l'autre quand λ est petit. On peut exploiter à nouveau la structure de gradient de la dynamique pour démontrer la même chose pour les directions de Y. De plus, l'ordre en λ peut être amélioré à $O(\sqrt{\lambda})$. La borne finale sur le biais que l'on obtient est la suivante :

$$W(P^\lambda, m_*\otimes\delta_{m_*}) = O(\sqrt{\lambda}),$$

où W désigne une distance de Wasserstein entre les projections de dimension finie des mesures de dimension infinie. Cette borne est également optimale en fonction de λ, comme cela peut être vérifié par un exemple gaussien.

Pour résumer, une plus petite valeur de λ conduit à un taux de convergence plus faible, mais réduit le biais de l'échantillonnage, confirmant les intuitions de nos discussions précédentes. Cependant, il convient de noter que le taux de convergence obtenu par le couplage par réflexion se détériore de façon exponentielle lorsque $\lambda \to 0$, rendant ce taux inadapté à l'analyse des dynamiques d'annealing en pratique.

Aperçu du chapitre 8

Dynamique de Schrödinger de champ moyen

Dans le dernier chapitre de la thèse, nous étudions le problème d'optimisation de champ moyen régularisé par l'information de Fisher (12) et le flot de gradient associé. Comme nous l'avons mentionné ci-dessus, le flot de gradient devrait au moins être la limite continue formelle du schéma discret de JKO (13). Par calcul des variations, on se rend compte que le flot discret est en fait l'Euler rétrograde :

$$\frac{\delta\mathfrak{F}}{\delta m}(\nu^h_{n+1},\cdot) + \frac{1}{h}\log\frac{\nu^h_{n+1}}{\nu^h_n} = \text{constante},$$

et on s'attend à ce que $\nu^h_{\lfloor t/h\rfloor}$ converge vers le flot m_t résolvant

$$\partial_t m_t = -\frac{\delta\mathfrak{F}}{\delta m}(m_t,\cdot)m_t + \lambda_t m_t,$$

où λ_t est la constante de normalisation

$$\lambda_t \coloneqq \int_{\mathcal{X}} \frac{\delta \mathfrak{F}}{\delta m}(m_t, x) m_t(\mathrm{d}x)$$

assurant que la masse est conservée : $\mathrm{d} \int_{\mathcal{X}} m_t / \mathrm{d}t = 0$. Rappelons que la fonctionnelle $\mathfrak{F}$ est régularisée par l'information de Fisher :

$$\mathfrak{F}(m) = F(m) + \int_{\mathcal{X}} \frac{|\nabla m|^2}{m}.$$

Par intégration par parties, on obtient l'expression suivante pour sa dérivée fonctionnelle linéaire :

$$\frac{\delta \mathfrak{F}}{\delta m}(m, x) = \frac{\delta F}{\delta m}(m, x) - 2\nabla \cdot \left(\frac{\nabla m}{m} \right) - \frac{|\nabla m|^2}{m^2}.$$

À ce stade, on peut déjà s'attendre à ce que le flot non linéaire m_t, une fois bien défini, converge vers la solution du problème d'optimisation (12) pour les deux raisons suivantes. Tout d'abord, par la formule ci-dessus, l'énergie régularisée $\mathfrak{F}$ cesse de diminuer seulement si $\frac{\delta \mathfrak{F}}{\delta m}(m_t, \cdot) - \lambda_t = 0$, c'est-à-dire que la mesure m_t est un point stationnaire du problème d'optimisation (12). Deuxièmement, le problème d'optimisation, étant la somme d'un F convexe au sens plat (qui est notre condition structurelle de base) et d'une information de Fisher strictement convexe au sens plat, n'a qu'un seul point stationnaire, et ce point est le minimiseur global. Compte tenu de ces intuitions, on peut démontrer rigoureusement la convergence par la compacité et le principe d'invariance de LaSalle, comme cela a été fait dans [117].

La question restante est de trouver un taux de convergence explicite et l'inégalité fonctionnelle derrière ce taux. Dans le cas de la Langevin de champ moyen sousamortie, la réponse est une inégalité de log-Sobolev uniforme comme le montrent [178, 56]. Et pour notre flot de gradient de champ moyen et de Fisher, nous avons besoin d'un trou spectral uniforme, ou en d'autres termes, d'une inégalité de Poincaré uniforme. Pour le voir, on effectue le changement de variable

$$\psi_t \coloneqq \sqrt{m_t}$$

et écrit l'équation dynamique pour ψ_t :

$$\partial_t \psi_t = 2\Delta \psi_t - \frac{1}{2} \frac{\delta F}{\delta m}(\psi_t^2, \cdot) \psi_t + \frac{1}{2} \lambda_t \psi_t.$$

Maintenant λ_t satisfait

$$\lambda_t = \int_{\mathcal{X}} 4|\nabla \psi_t|^2 + \frac{\delta F}{\delta m}(\psi_t, \cdot) \psi_t^2$$

et est la constante garantissant que ψ_t est normalisé dans L^2. Dans le cas linéaire, la dérivée plate ne dépend pas de la mesure :

$$\frac{\delta F}{\delta m}(m, x) = U(x),$$

et l'évolution de ψ_t correspond à un semi-groupe de Schrödinger linéaire. La convergence exponentielle est ainsi garantie par le trou spectral de l'opérateur hamiltonien :

$$\mathcal{H} = -4\Delta + U.$$

En revenant à l'évolution non linéaire, on définit le hamiltonien à chaque instant :

$$\mathcal{H}_t \coloneqq -4\Delta + \frac{\delta F}{\delta m}(m_t, \cdot).$$

Alors

$$\partial_t \psi_t = -\frac{1}{2}(\mathcal{H}_t - \lambda_t)\psi_t,$$
$$\lambda_t = (\psi_t, \mathcal{H}_t \psi_t)_{L^2}.$$

La diminution de $\mathfrak{F}(m_t)$ satisfait

$$\frac{\mathrm{d}\mathfrak{F}(m_t)}{\mathrm{d}t} = -\big(\psi_t, (\mathcal{H}_t - \lambda_t)\mathcal{H}_t\psi_t\big)_{L^2} = -(\psi_t, \mathcal{H}_t^2\psi_t)_{L^2} + (\psi_t, \mathcal{H}_t\psi_t)^2_{L^2}.$$

Notons $\hat{\psi}_t$ l'état fondamental normalisé unique de $\mathcal{H}_t$. On obtient, par le trou spectral,

$$(\psi_t, \mathcal{H}_t^2\psi_t)_{L^2} - (\psi_t, \mathcal{H}_t\psi_t)^2_{L^2} \gtrsim (\psi_t, \mathcal{H}_t\psi_t)_{L^2} - (\hat{\psi}_t, \mathcal{H}_t\hat{\psi}_t)_{L^2}.$$

De nouveau, en revenant aux variables de mesure et en utilisant la convexité de F, on peut déduire que

$$(\psi_t, \mathcal{H}_t\psi_t)_{L^2} - (\hat{\psi}_t, \mathcal{H}_t\hat{\psi}_t)_{L^2} \geqslant \mathfrak{F}(m_t) - \inf \mathfrak{F}.$$

On a donc la convergence exponentielle :

$$\mathfrak{F}(m_t) - \inf \mathfrak{F} \leqslant Ce^{-ct},$$

étant donné le trou spectral uniforme pour $\mathcal{H}_t$. Il est bien connu que le trou spectral uniforme est équivalent à une inégalité de Poincaré uniforme pour la mesure de probabilité $\hat{m}_t \coloneqq \hat{\psi}_t^2$ résolvant l'équation stationnaire

$$\frac{\delta F}{\delta m}(m_t, x) - 2\nabla \cdot \left(\frac{\nabla \hat{m}_t}{\hat{m}_t}\right) - \frac{|\nabla \hat{m}_t|^2}{\hat{m}_t^2} = \text{constante}.$$

En notant la densité logarithmique par $\hat{u}_t \coloneqq -\log \hat{m}_t$, on trouve que $\hat{u}_t$ résout l'équation de HJB ergodique

$$2\Delta \hat{u}_t - |\nabla \hat{u}_t|^2 + \frac{\delta F}{\delta m}(m_t, x) = \text{constante}.$$

Sous l'hypothèse que $\frac{\delta F}{\delta m}(m, \cdot)$ est une somme d'une fonction fortement convexe et d'une fonction lipschitzienne, uniformément en m, on peut utiliser la méthode de Conforti [61] pour obtenir que $\hat{u}_t$ est également une somme d'une partie fortement convexe et d'une partie lipschitzienne avec des bornes uniformes. Une inégalité de Poincaré uniforme suit alors de par exemple [9].

Avancées récentes et perspectives

Un inconvénient commun des chapitres 1 et 2, comme l'a souligné un relecteur anonyme, est que nous ne comparons *pas directement* le système de particules m_t^N et le flot de champ moyen m_t à long terme. Au lieu de cela, cette comparaison est faite via la mesure invariante de champ moyen m_*, complétée par une borne globale en temps. Cet argument triangulaire est plutôt maladroit et entraîne une perte d'exposant dans la borne finale de propagation du chaos. Nous annonçons que nous résoudrons ce problème par une méthode de comparaison directe, où nous travaillons avec une distance entre les mesures de probabilité qui est induite par le paysage d'énergie libre, et retrouvons l'ordre d'erreur optimale $O(1)$ (ou $O(1/N)$, selon la normalisation). Nous explorerons également d'autres conséquences de l'ISL non linéaire (17) et de sa version à N particules (16), telles que la concentration de la mesure uniforme en temps pour le système de particules de Langevin de champ moyen et les propriétés de turnpike pour le problème de Schrödinger de champ moyen associé.

Dans un travail récent de l'auteur [230], l'ISL défectueuse (16) établie au chapitre 1 a été tendue en une ISL N-uniforme grâce à l'utilisation d'une inégalité de Poincaré supplémentaire. Cette approche offre une alternative au travail simultané de Chewi, Nitanda et M. S. Zhang [55], tout en fournissant une meilleure dépendance à la force d'interaction du champ moyen.

Plus récemment, Bauerschmidt, Bodineau et Dagallier [14] ont adapté la méthode du flot de Polchinski aux systèmes de particules à champ moyen et ont établi une ISL N-uniforme sur l'ensemble du régime d'unicité. Plus précisément, le fonctionnel d'énergie libre $\mathcal{F}$ est autorisé à inclure une composante d'énergie *concave* au sens plat, et l'analyse est menée directement sous une forme projetée de l'ISL non linéaire :

$$I(m|\hat{m}) \gtrsim H(m|\hat{m}) \gtrsim \mathcal{F}(m) - \mathcal{F}(m_*).$$

Cette hypothèse est plus faible que la convexité plate et permet de retrouver le comportement critique de Curie-Weiss. Cependant, la méthode semble moins intrinsèque pour les interactions non quadratiques et conduit à une constante d'ISL plus faible. Nous annonçons ici que, dans un travail à venir, nous établirons une ISL défectueuse N-uniforme via une approche intrinsèque qui dépend uniquement du paysage d'énergie libre non projeté. De plus, notre méthode correspond à un schéma de *localisation stochastique par coordonnées*, tandis que la leur est un schéma de *bascule linéaire* dans le langage de Y. Chen-Eldan [52].

Dans un autre travail récent [194], nous étudions la question de la taille du chaos pour la dynamique de Langevin suramortie sous la condition d'ISL non linéaire mentionnée ci-dessus. Plus précisément, nous montrons que

$$H(m_*^{N,k} \mid m_*^{\otimes k}) = O\left(\frac{k^2}{N^2}\right),$$

où $m_*^{N,k}$ désigne la distribution k-marginale de la mesure de Gibbs à N particules. Dans ce travail, nous identifions une structure de gradient pour les mesures conditionnelles et développons une hiérarchie d'entropie d'un ordre supérieur à la formulation originale de Lacker. Cette approche non perturbative de l'interaction de champ moyen étend la littérature existante [141, 142, 20], qui ne traite que des scénarios où l'interaction est effectivement dominée par la diffusion. Néanmoins, le problème dynamique du chaos uniforme en temps reste en grande partie ouvert et mérite clairement des recherches supplémentaires.

Pour le modèle singulier de vortex en 2D, le problème de la taille du chaos n'est pas complètement résolu dans la thèse actuelle car notre méthode échoue dans le régime de basse température. La résolution complète de ce problème nécessite une étude supplémentaire, mais il semble à l'auteur que certains éléments cruciaux font encore défaut. De plus, nous pouvons également considérer le problème de la taille du chaos pour les interactions de Coulomb ou de Riesz en dimension supérieur. Cela semble encore plus difficile à l'auteur en raison de la singularité plus forte dans le noyau d'interaction.

L'étude des systèmes de Vlasov-Poisson a récemment connu des avancées significatives, plusieurs idées et techniques novatrices pour établir la propagation du chaos ayant été introduites dans [30, 29, 51]. Toutefois, le cas non régularisé en dimension $\geqslant 3$, aussi bien dans le cadre diffusif que non diffusif, demeure un problème ouvert.

On peut également se demander si l'estimation entropique elliptique cruciale du chapitre 7 peut être étendue au cas dynamique parabolique. Si cela réussit, une telle approche fournirait des propriétés de contractivité plus fortes que celles obtenues par la méthode de couplage. Nous avons également l'intention d'étudier la dynamique cinétique auto-interagissant et d'explorer l'utilisation de l'auto-interaction dans le cadre des jeux à champ moyen.

Alors que les études sur la dynamique de Schrödinger de champ moyen dans le chapitre 8 se concentrent sur la partie théorique, il est tout aussi important d'explorer ses aspects numériques et son efficacité dans les applications réelles.

*
* *

Les huit chapitres de la thèse sont d'abord parus individuellement sous forme de publications [50, 48, 171, 170, 49] et de prépublications [229, 77, 60]. Pour cette raison, les notations et conventions dans les différents chapitres peuvent ne pas être cohérentes. Elles peuvent également être différentes de celles utilisées dans cette introduction.

Part I

Overdamped and underdamped mean field Langevin dynamics

Chapter 1

Uniform-in-time propagation of chaos for mean field Langevin dynamics

Abstract. We study the mean field Langevin dynamics and the associated particle system. By assuming the functional convexity of the energy, we obtain the L^p-convergence of the marginal distributions toward the unique invariant measure for the mean field dynamics. Furthermore, we prove the uniform-in-time propagation of chaos in both the L^2-Wasserstein metric and relative entropy.

Based on joint work with Fan Chen and Zhenjie Ren.

1.1 Introduction

1.1.1 Preview of main results

Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a *mean field functional* and $D_m F$ be its intrinsic derivative. In this paper, we study the long-time behavior of the following mean field Langevin (MFL) dynamics:

$$\mathrm{d}X_t = -D_m F(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \qquad \text{where } m_t = \mathrm{Law}(X_t), \tag{1.1}$$

as well as the corresponding dynamics of N particles:

$$\mathrm{d}X_t^i = -D_m F(\mu_{\mathbf{X}_t}, X_t^i)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i, \quad i = 1, \ldots, N, \quad \text{where } \mu_{\mathbf{X}_t} = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i}.$$

Here, W_t, W_t^i are independent d-dimensional standard Brownian motions. We suppose that F is a functional such that

- the mapping $m \mapsto F(m)$ is convex in the functional sense (as opposed to the optimal transport sense);
- for every $x \in \mathbb{R}^d$, the mapping $m \mapsto D_m F(m, x)$ is M_{mm}^F-Lipschitz continuous with respect to the L^1-Wasserstein metric;

- for every $m \in \mathcal{P}_2(\mathbb{R}^d)$, the probability measure on $\mathbb{R}^d$ that has density proportional to $x \mapsto \exp\big(-\frac{\delta F}{\delta m}(m,x)\big)$ satisfies the ρ-logarithmic Sobolev inequality (LSI) for some $\rho > 0$.

Recently, there has been a growing interest in modeling the training of neural networks as a convex mean field optimization problem (see [163, 57, 211, 203, 117, 128, 63] and also our Section 1.3 for explanations). With some exceptions (e.g., [57, 176, 179] and Chapters 6 and 8), the majority of the studies [163, 117, 56, 178] have focused on the entropy-regularized mean field optimization problem and the corresponding MFL dynamics in the form of (1.1). It was first proved in [117] that under the convexity assumption of F, the marginal distributions of the MFL dynamics converge toward its unique invariant measure, which is also the unique minimizer of the mean field optimization problem. Then it is shown in [178, 56] that, with the presence of the uniform LSI, such kind of convergence is exponentially fast. The main contribution of this paper lies in that, we further explore the fine properties of MFL dynamics with a particular emphasis on its uniform-in-time propagation of chaos property, i.e., the time-uniform upper bounds for the distance between the finite-particle and the mean field dynamics. Therefore, we provide a theoretical guarantee for the applicability of the finite-particle approximation when the dynamics is expected to run for an indefinitely long time.

Recall that we have defined $m_t = \operatorname{Law}(X_t)$. Let us also define

$$m_t^N = \operatorname{Law}(X_t^1, \dots, X_t^N)$$

and denote by m_∞ the unique invariant measure of the mean field dynamics. Our main results are summarized as follows:

- if the Radon–Nikodým derivative $\mathrm{d}m_0/\mathrm{d}m_\infty$ belongs to $L^{p_0}(m_\infty)$ for some $p_0 > 1$, then for every $p \in \mathbb{R}$, the norm $\|\mathrm{d}m_t/\mathrm{d}m_\infty\|_{L^p(m_\infty)} \to 1$ exponentially fast when $t \to \infty$;

- the scaled L^2-Wasserstein distance and the relative entropy $\frac{1}{N}W_2^2(m_t^N, m_\infty^{\otimes N})$, $\frac{1}{N}H(m_t^N|m_t^{\otimes N})$ converge to a $O(N^{-1})$ neighborhood of zero when $t \to \infty$, with an exponential rate that is independent of N;

- if the initial error is zero, i.e., $m_0^N = m_0^{\otimes N}$, then

  $$\sup_{t\in[0,\infty)} \frac{1}{N} W_2^2(m_t^N, m_t^{\otimes N}) \to 0$$

  when $N \to \infty$; further if the assumption of the first claim holds, then

  $$\sup_{t\in[0,\infty)} \frac{1}{N} H(m_t^N|m_t^{\otimes N}) \to 0$$

  when $N \to \infty$.

We also refer those interested readers to Chapter 2, which delves into analogous properties for kinetic MFL dynamics.

1.1.2 Related works

Long-time behavior of McKean–Vlasov dynamics. Propagation of chaos in finite time for the stochastic McKean–Vlasov dynamics

$$\mathrm{d}X_t = b(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \qquad \text{where } m_t = \mathrm{Law}(X_t)$$

is relatively easy to show, using the *synchronous coupling* approach, given that b is a jointly Lipschitz function of both measure and space variables in the sense of the Wasserstein metric. The bound obtained by this method, however, generally tends to infinity when the time interval extends to infinity. Besides, the dynamics may possess multiple invariant measures, so uniform-in-time convergence can not be expected without some additional assumptions or a more general definition of convergence itself (e.g. convergence modulo symmetries).

The research on the long-time behavior of McKean–Vlasov dynamics has been active in recent years and here we introduce a setting that has appeared in many previous works. Consider functions U, $V : \mathbb{R}^d \to \mathbb{R}$ and the following special kind of drift

$$b(m, x) = -\nabla U(x) - \int \nabla V(x - \tilde{x}) m(\mathrm{d}\tilde{x}).$$

In this case, U is referred as the *external potential* and V is called the *interaction potential*.

In this paragraph, we provide a far from exhaustive review of uniform-in-time propagation of chaos (POC) for McKean–Vlasov dynamics. First, in the work [159] of Malrieu in 2001, uniform POC is established by synchronous coupling for overdamped dynamics under the assumption that U is strongly convex and V is convex. In an alternative way, Carrillo, McCann and Villani set up the mean field gradient flow framework in their work [39], which our paper also relies on. They showed the exponential convergence of the overdamped mean field system under the assumption that $U + 2V$ is strongly convex. In Monmarché's work [167], uniform POC is extended to the *kinetic* Langevin dynamics, assuming the same convexity assumption on $U + 2V$. This assumption is further relaxed in his follow-up work with Guillin [101], where they incorporate the uniform-in-N log-Sobolev inequality in [100]. In [80], Durmus, Eberle, Guillin and Zimmer showed uniform POC for overdamped Langevin dynamics, under the assumption that the confining potential U is only weakly convex and V is small enough, utilizing a *reflection coupling* technique. The reflection coupling technique is then used by Schuh in [206] to show uniform POC for kinetic Langevin dynamics, albeit in this setting, the form of the confining potential is more restricted compared to the overdamped case. The weak uniform-in-time convergence is also demonstrated for the overdamped dynamics on a torus in [70] by Delarue and Tse under various settings. This research assumes the smallness of interaction without explicitly specifying its form and employs a *master equation* analysis. In [142], Lacker and Le Flem showed a sharp $O(1/N^2)$ rate for time-uniform propagation of chaos for the overdamped dynamics, by studying the relative entropy growth between marginal distributions with the help of a time-uniform log-Sobolev inequality for the mean field flow.

We now comment on the assumptions and methods of these works. Apart from the second and third settings of [70] and that of [142], the aforementioned works all rely on the smallness or the (semi-, weak) convexity of the interaction potential. This smallness or convexity is used to control the error between the coupled processes, or to deduce a uniform-in-N log-Sobolev inequality for the N-particle system's invariant measure (see [100]). Our setting is different from those in other works. First, our results are built upon the functional convexity of the mean field energy functional, which is a different (and even exclusive in some cases) assumption from the convexity of the interaction potential. Further details on this alternative assumption of convexity will be provided in the following paragraph. Second, our approach does not rely on a uniform-in-N log-Sobolev inequality for the invariant measure of the N-particle system.

Finally, we remark that the translation-invariant models have been studied in the last setting of [70] and also in [79]. In these cases, there exists a continuum of invariant measures, and the POC is then obtained modulo the translational symmetry. Besides, we also mention that in a recent work [98], Guillin et al. studied the 2D viscous vortex model where the particles are in singular interactions and showed the uniform POC estimates.

Linear functional convexity. One of our key assumptions is the (linear functional) convexity of the mean field functional F, formally defined in (1.2). Except in [215, 70], this assumption has not been explicitly exploited to investigate the long-time behavior of the McKean–Vlasov dynamics. It is important to distinguish this convexity from the displacement convexity, which frequently appears in the optimal transport literature and is defined in, for instance, [222, Definition 16.1]. We will clarify in Remark 1.18 that, for continuous two-body interaction potentials, Bochner's theorem implies that these two concepts are even mutually exclusive, except in trivial cases.

This particular form of convexity is implicitly exploited in [70] to obtain time-uniform POC estimates. More precisely, the authors studied McKean–Vlasov drift of form $b(m,x) = -\int \nabla V(x-\tilde{x})m(\mathrm{d}\tilde{x})$ on the torus, where all Fourier coefficients of the interaction potential V are nonnegative. Then this property is used to obtain estimates on the master equation in the long time. We note that, here, the positivity of the Fourier coefficient implies that the corresponding energy $F(m) = \frac{1}{2}\iint V(x-\tilde{x})m(\mathrm{d}x)m(\mathrm{d}\tilde{x})$ is convex in our functional sense. Although our results are stated for dynamics in $\mathbb{R}^d$, it is reasonable to expect that our methodology can be extended to the torus and yield similar results.

The primary motivation for introducing this new setting is to study the training of two-layer (or one-hidden-layer) neural networks, which we will explain in Examples 1.21 and 1.27.

Gradient descent. Our dynamics is a special case of McKean–Vlasov with the drift of gradient type:

$$b(m,x) = -D_m F(m,x) = -\nabla \frac{\delta F}{\delta m}(m,x).$$

This form of drift corresponds to the gradient descent of the *free energy* $\mathcal{F} = F + H$ in L^2-Wasserstein space, here, $H(m) = \int m(x)\log m(x)\,\mathrm{d}x$ is the (absolute) entropy of the measure. We refer the readers to [126] for detailed discussions about the gradient flow with the linear energy $F(m) = \int V(x)m(\mathrm{d}x)$, and [4] for a general gradient flow framework in Wasserstein space. We note that, in a previous work [117], this gradient flow structure is exploited to obtain the ergodicity of the MFL dynamics. Precisely, the authors established the following free energy dissipation formula

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\int |D_m F(m_t, x) + \nabla \log m_t(x)|^2 m_t(\mathrm{d}x),$$

and then by combining this with LaSalle's invariance principle and the uniqueness of the invariant measure, they showed the global convergence of the MFL dynamics. In this paper, we will prove the same energy descent formula under weaker assumptions on the regularity of $x \mapsto D_m F(m, x)$, thanks to the general framework developed in [4].

1.1.3 Main contributions

L^p convergence and hypercontractivity of MFL. The exponential convergence of relative entropy for the MFL with convex F has been proved in [56, 178] via log-Sobolev inequalities, extending the classical result [180] wherein the F is linear in measure. In this paper, we introduce a stronger L^p-convergence in Theorem 1.9. To achieve this enhanced convergence result, we require the initial condition to lie in L^{p_0} for some $p_0 > 1$. This contrasts with the situation of relative entropy, where elliptic regularization ensures relative entropy to be finite at all positive times (see Proposition 1.37).

Our method of proof is based on the L^2-convergence and the hypercontractivity, which ports the L^2-convergence to L^p for all $p \in \mathbb{R}$. Two pivotal observations are the growth of L^p-norm formula (1.51) and the hypercontractive inequalities (1.20), (1.21) for the mean field flow. Recently the hypercontractivity has also been ultilized in [59] to show the L^p-convergence of MFL with Riesz interactions (though on a torus). Finally, it is important to mention that the proof of our propagation of chaos result (Theorem 1.14) requires the L^p-convergence for p negative. To address this requirement, we establish the *reverse* hypercontractivity of the MFL. This property follows from the analogous formal computations to those employed in direct hypercontractivity, under the assumption that the invariant measure satisfies a LSI.

Convergence of particle system. Within the mean field setting established in [56, 178], we show in Theorem 1.12 that the particle system's free energy converges to the N-tensorized invariant measure of the mean field system exponentially modulo an error of size $O(N^{-1})$ per particle. Our proof approach relies on a decomposition of relative Fisher information and a componentwise application of the log-Sobolev inequality, which introduces the $O(N^{-1})$ error per particle. Our result differs from that of [100], where the precise convergence of the particle system to its invariant measure is obtained through the use of the uniform-in-N log-Sobolev inequality. One notable advantage of our method is that we allow applications involving potentially significant interactions, including cases such as the training of neural networks (as discussed in Examples 1.21 and 1.27.)

Propagation of chaos. By combining the two previous results, i.e. the L^p-convergence of the MFL and the entropic convergence of the particle system, we are able to control the distance between the particle system m_t^N and N-tensorized mean field flow $m_t^{\otimes N}$, in terms of Wasserstein distance *and* relative entropy. The bound on Wasserstein is a direct consequence of Talagrand's T_2 transport inequality. To control the relative entropy we employ a classical duality formula (1.55) to link $H(m_t^N|m_t^{\otimes N})$ to the $-p$ norm $\|\,\mathrm{d}m_t/\mathrm{d}m_\infty\|_{-p}$ for $p > 0$, whose exponential convergence is guaranteed by Theorem 1.9. As a side result, we also obtain the uniform-in-time concentration of measure of the mean field flow (Theorem 1.11), based on this observation.

Let us now compare our method to those of [142, 215]. In [142] the authors assumed the mean field flow satisfies a uniform LSI and utilized an entropy growth formula similar to our L^p-growth formula to estimate the relative entropy bound. As remarked in [215], verifying this uniform LSI can be challenging in the mean field setting. In particular if one wishes to apply the Holley–Stroock perturbation lemma to the invariant measure m_∞, the mean field flow needs to satisfy $\log \mathrm{d}m_t/\mathrm{d}m_\infty \in L^\infty$ uniformly In [215], Suzuki, Nitanda and Wu made the assumptions that the confining potential exhibits a super-quadratic growth, so that this boundedness follows from the ultracontractivity via super LSI. However, this confining potential is stronger than the quadratic one in our setting and the constants derived from ultracontractivity are dependent on the spatial dimension.

1.1.4 Notations

Let d be a positive integer and x an element of $\mathbb{R}^d$. We denote the Euclidean norm of $x \in \mathbb{R}^d$ by $|x|$ and define c_d as the volume of the d-dimensional unit ball. Let $p \geqslant 1$, we define $\mathcal{P}_p(\mathbb{R}^d)$ to be the space of probability measures on $\mathbb{R}^d$ with finite p-moment, i.e., $\mathcal{P}_p(\mathbb{R}^d) = \{m \in \mathcal{P}(\mathbb{R}^d) : \int |x|^p m(\mathrm{d}x) < +\infty\}$. The L^p-Wasserstein metric is denoted by W_p and its definition along with elementary properties, can be found in [4, Chapter 7].

Consider a mean field functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. We denote by $\frac{\delta F}{\delta m} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ its linear functional derivative and by $D_m F = \nabla \frac{\delta F}{\delta m} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ its intrinsic derivative, provided they exist. The definition of linear functional derivative on $\mathcal{P}_2(\mathbb{R}^d)$ can be found in [37, Definition 5.43].

Let X, Y be two random variables. We denote the distribution of X as $\mathrm{Law}(X)$ and write $X \sim m$ when $m = \mathrm{Law}(X)$. Additionally, we use $X \overset{d}{=} Y$ to indicate that $\mathrm{Law}(X) = \mathrm{Law}(Y)$. The set of couplings between probability measures μ, ν is denoted by $\Pi(\mu, \nu)$. Let $N \geqslant 2$ be an integer, we use the bold letter $\boldsymbol{x}_N = (x^1, \ldots, x^N)$ to represent an N-tuple of the elements in $\mathbb{R}^d$. We omit the subscript N when there are no ambiguities.

Let $I \subset \{1, \ldots, N\}$. We define $-I \coloneqq \{1, \ldots, N\} \setminus I$, i.e., the complementary index set of I. For a probability measure $m^N = \mathrm{Law}(\boldsymbol{X}) \in \mathcal{P}(\mathbb{R}^{dN})$, we denote its marginal and the (regular) conditional distributions by

$$
\begin{aligned}
m^{N,I} &= \mathrm{Law}(X^i)_{i \in I}, \\
m^{N,I|-I}(\boldsymbol{x}^{-I}) &= \mathrm{Law}\big((X^i)_{i \in I}\big|X^j = x^j,\ j \in -I\big),
\end{aligned}
$$

where the latter is defined $m^{N,-I}$-almost surely and $\boldsymbol{x}^{-I}$ denotes the tuple $(x^j)_{j \in -I}$. We identify i with the singleton $\{i\}$ when working with indices.

Given $\boldsymbol{x}_N = (x^1, \ldots, x^N) \in \mathbb{R}^{dN}$, we denote the corresponding empirical measure by

$$
\mu_{\boldsymbol{x}_N} = \frac{1}{N} \sum_{i=1}^N \delta_{x^i}.
$$

For $i = 1, \ldots, N$, as introduced in the paragraph above, the symbol $-i$ denotes the complementary set $\{1, \ldots, N\} \setminus i$. We denote the empirical measure of the $N-1$

points $\boldsymbol{x}_N^{-i} = (x_j)_{j\neq i}$ by

$$\mu_{\boldsymbol{x}_N^{-i}} = \frac{1}{N-1}\sum_{j=1,\, j\neq i}^{N} \delta_{x^j}.$$

For a $\mathbb{R}^{dN}$-valued random variable $\boldsymbol{X}_N = (X^i)_{i=1}^N$, we can thereby form the random empirical measures $\mu_{\boldsymbol{X}_N}$, $\mu_{\boldsymbol{X}_N^{-i}}$.

When a measure $m \in \mathcal{P}(\mathbb{R}^d)$ has a density with respect to the d-dimensional Lebesgue measure, we still denote its density function by $m : \mathbb{R}^d \to \mathbb{R}$. Let γ be a positive and σ-finite measure on $\mathbb{R}^d$. We define the relative entropy

$$H(m|\gamma) = \int \log \frac{\mathrm{d}m}{\mathrm{d}\gamma}(x) m(\mathrm{d}x)$$

and the relative Fisher information

$$I(m|\gamma) = \int \left|\nabla \log \frac{\mathrm{d}m}{\mathrm{d}\gamma}\right|^2 m(\mathrm{d}x)$$

provided the corresponding integrals are well defined. In cases where the integrals are not well defined, we set H, $I = +\infty$ respectively. When $\gamma = \mathcal{L}^d$ is the Lebesgue measure on $\mathbb{R}^d$, we omit the dependence on γ and define the *absolute* entropy and Fisher information as:

$$H(m) \coloneqq H(m|\mathcal{L}^d), \qquad I(m) \coloneqq I(m|\mathcal{L}^d),$$

provided they are well-defined. For nonnegative functions $f : \mathbb{R}^d \to [0, +\infty)$ we also define its entropy as

$$\mathrm{Ent}_m f = \mathbb{E}_m[f \log f] - \mathbb{E}_m[f] \log \mathbb{E}_m[f],$$

which is well defined in $[0, +\infty]$ according to Jensen's inequality.

Organization of paper. In Section 1.2, we present our assumptions, introduce the mean field Langevin dynamics and the particle system, and state our main results. In section 1.3, we offer some examples of MFL, to which our theorems can be applied, accompanied by numerical experiments of two-layer neural network training. The proofs are given in the rest of the paper, and for the most technically demanding ones, we detail them in Appendix A.1. We also show a modified Bochner's theorem in Appendix A.2.

1.2 Main results

Assumptions. Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a mean field functional. We suppose F is convex in the sense that for all $t \in [0, 1]$ and all m, $m' \in \mathcal{P}_2(\mathbb{R}^d)$,

$$F\big((1-t)m + tm'\big) \leqslant (1-t)F(m) + tF(m'). \tag{1.2}$$

Suppose also its intrinsic derivative $D_m F : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ exists and satisfies

$$\forall x \in \mathbb{R}^d,\ \forall m, m' \in \mathcal{P}_2(\mathbb{R}^d),\ |D_m F(m, x) - D_m F(m', x)| \leqslant M_{mm}^F W_1(m, m') \tag{1.3}$$

for some constant $M^F_{mm} \geqslant 0$. For each $m \in \mathcal{P}_2(\mathbb{R}^d)$, we define a probability measure $\hat{m}$ by its density

$$\hat{m}(x) \propto \exp\Big(-\frac{\delta F}{\delta m}(m,x)\Big)$$

and suppose $\hat{m}$ satisfies the *ρ-logarithmic Sobolev inequality* (LSI) uniformly in m for some $\rho > 0$, that is, for every $m \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\forall f \in C^1_b(\mathbb{R}^d), \qquad \rho \operatorname{Ent}_{\hat{m}}(f^2) \leqslant \mathbb{E}_{\hat{m}}[|\nabla f|^2]. \tag{1.4}$$

Here, we implicitly suppose that $\hat{m}$ is well defined for all $m \in \mathcal{P}_2(\mathbb{R}^d)$, and in particular, we have $\int \exp\big(-\frac{\delta F}{\delta m}(m,x)\big)\,\mathrm{d}x < \infty$. We remark that the inequality above can be verified for mean field functionals F whose linear derivative $\frac{\delta F}{\delta m}$ is a perturbation of a strongly convex function. For details, we refer readers to Proposition 1.25 in Section 1.3.2. We suppose as well

$$\sup_{m\in\mathcal{P}_2(\mathbb{R}^d)} \sup_{x\in\mathbb{R}^d} |\nabla D_m F(m,x)| \leqslant M^F_{mx} \tag{1.5}$$

for some constant $M^F_{mx} \geqslant 0$. Finally, for some of the results we additionally suppose that $x \mapsto D_m F(m,x)$ belongs to C^3 with the bounds

$$\sup_{m\in\mathcal{P}_2(\mathbb{R}^d)} \sup_{x\in\mathbb{R}^d} |\nabla^k D_m F(m,x)| < +\infty, \qquad k = 2,\ 3. \tag{1.6}$$

Remark 1.1 (Well-definedness of $\hat{m}$). The definition of $\hat{m}$ relies on the finiteness of the normalization constant

$$Z(\hat{m}) = \int \exp\Big(-\frac{\delta F}{\delta m}(m,x)\Big)\,\mathrm{d}x. \tag{1.7}$$

As mentioned above, it is assumed implicitly in the condition (1.4) that $Z(\hat{m})$ is finite for every $m \in \mathcal{P}_2(\mathbb{R}^d)$. We will prove in Proposition 1.32 that the following is sufficient for this finiteness:

- the condition (1.3) holds, and
- there exists at least one measure m_0 such that $Z(\hat{m}_0)$ is finite and m_0 satisfies the LSI (1.4).

Remark 1.2 (Functional inequalities). By approximating the function f by a sequence of functions in C^1_{b}, we find that the inequality (1.4) holds for functions f whose generalized derivative satisfies $\mathbb{E}_{\hat{m}}[|\nabla f|^2] < +\infty$. It is well known that the LSI (1.4) implies the *Poincaré inequality*:

$$\forall f \in C^1_b(\mathbb{R}^d), \qquad 2\rho \operatorname{Var}_{\hat{m}}(f) \leqslant \mathbb{E}_{\hat{m}}[|\nabla f|^2]. \tag{1.8}$$

The restriction $f \in C^1_{\mathrm{b}}$ can be analogously removed. The LSI (1.4) also implies *Talagrand's T_2-transport inequality*:

$$\forall \mu \in \mathcal{P}_2(\mathbb{R}^d), \qquad \rho W_2^2(\mu,\hat{m}) \leqslant H(\mu|\hat{m}). \tag{1.9}$$

See the original work of Otto and Villani [180] for a proof. We also sketch their argument in the proof of Lemma 1.31. All those three inequalities, namely (1.4), (1.8), (1.9), are stable under tensorization: if one replaces, for some $N \geqslant 2$, the measure $\hat{m}$ by its tensor product $\hat{m}^{\otimes N}$, which is a measure on $\mathbb{R}^{dN}$, and the function $f : \mathbb{R}^d \to \mathbb{R}$ (resp. the probability measure μ on $\mathbb{R}^d$) by function $f^N : \mathbb{R}^{dN} \to \mathbb{R}$ having a square-integrable weak derivative ∇f^N with respect to the measure $\hat{m}^{\otimes N}$ (resp. probability measures μ^N on $\mathbb{R}^{dN}$), then the inequalities hold with the same constant ρ.

Mean field and particle system. We study the *mean field Langevin dynamics*, that is, the following McKean–Vlasov SDE

$$\mathrm{d}X_t = -D_m F(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \qquad \text{where } \mathrm{Law}(X_t) = m_t. \tag{1.10}$$

Let $N \geqslant 2$. The corresponding N-*particle system* is defined by

$$\mathrm{d}X_t^i = -D_m F(\mu_{\boldsymbol{X}_t}, X_t^i)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i, \quad i = 1, \ldots, N, \quad \text{where } \mu_{\boldsymbol{X}_t} = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}. \tag{1.11}$$

Here, W, W^i are standard Brownian motions in $\mathbb{R}^d$, which are independent from each other. Their marginal distributions $m_t = \mathrm{Law}(X_t)$, $m_t^N = \mathrm{Law}(\boldsymbol{X}_t) = \mathrm{Law}(X_t^1, \ldots, X_t^N)$ then solve the Fokker–Planck equations respectively

$$\partial_t m_t = \Delta m_t + \nabla \cdot \big(D_m F(m_t, \cdot) m_t\big), \tag{1.12}$$

$$\partial_t m_t^N = \sum_{i=1}^N \Big(\Delta_i m_t^N + \nabla_i \cdot \big(D_m F(\mu_{\boldsymbol{x}}, x^i) m_t^N\big)\Big). \tag{1.13}$$

The mean field equation (1.12) is non-linear while the N-particle system equation (1.13) is linear. We will prove in Proposition 1.37 that, if the initial condition $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$, the mean field dynamics (1.12) is well posed and enjoys certain regularity.

Remark 1.3. We have fixed the volatility (diffusion) constant to be $\sqrt{2}$ to simplify our computations. In order to apply our results to the MFL defined by

$$\mathrm{d}X_t = -D_m F(m_t, X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \qquad \text{where } \mathrm{Law}(X_t) = m_t,$$

with some $\sigma > 0$, we apply the rescaling: $\tilde{t} = \frac{\sigma^2}{2} t$, $\tilde{F} = \frac{2}{\sigma^2} F$ and $\tilde{X}_{\tilde{t}} = X_t$. In this way, the new diffusion process $\tilde{t} \mapsto \tilde{X}_{\tilde{t}}$ satisfies the SDE (1.10), whose diffusion constant is fixed to $\sqrt{2}$, with the new mean field functional $\tilde{F}$. The same scaling transform can be applied to the particle system as well.

Free energy and invariant measure. We focus on the long-term behavior of the MFL (1.12) and the corresponding particle system (1.13), where invariant measures play a key role. Define *mean field free energy* $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ by

$$\mathcal{F}(m) = F(m) + H(m). \tag{1.14}$$

Given the assumptions (1.2), (1.3), (1.4) and (1.5), we can show the existence of a unique minimizer of $\mathcal{F}$, denoted by m_∞. Furthermore, this measure m_∞ satisfies the *first-order condition*:

$$m_\infty(\mathrm{d}x) = \hat{m}_\infty(\mathrm{d}x) = \frac{1}{Z(\hat{m}_\infty)} \exp\bigg(-\frac{\delta F}{\delta m}(m_\infty, x)\bigg)\,\mathrm{d}x. \tag{1.15}$$

The precise statement and proof is given in Proposition 1.34. Differentiating both sides of the first-order condition, we obtain $\Delta m_\infty + \nabla \cdot \big(D_m F(m_\infty, x) m_\infty\big) = 0$, which implies that m_∞ is an invariant measure to mean field Fokker–Planck equation (1.12). Conversely, we will show in Corollary 1.39 that under our conditions every invariant measure satisfies the first-order condition and, therefore, we get the uniqueness of invariant measure as well.

The N-particle system (1.11) is a classical Langevin dynamics because the equation (1.13) is linear. We define the *N-particle free energy* $\mathcal{F}^N : \mathcal{P}_2(\mathbb{R}^{dN}) \to (-\infty, +\infty]$ by

$$\mathcal{F}^N(m^N) = N \int F(\mu_{\boldsymbol{x}}) m^N(\mathrm{d}\boldsymbol{x}) + H(m^N). \tag{1.16}$$

We will prove in Proposition 1.33 that under our assumptions (1.2), (1.3) and (1.4), the minimizer m^N_∞ of $\mathcal{F}^N$ exists, and has the density

$$m^N_\infty(\mathrm{d}\boldsymbol{x}) \propto \exp\bigl(-NF(\mu_{\boldsymbol{x}})\bigr)\,\mathrm{d}\boldsymbol{x}, \tag{1.17}$$

which is invariant to the N-particle Fokker–Planck equation (1.13). By the definition of free energy we have $\mathcal{F}^N(m^N) = H(m^N|m^N_\infty) + \text{constant}$, so m^N_∞ also minimizes the N-particle free energy $\mathcal{F}^N$.

L^p_+ space for all $p \in \mathbb{R}$. We investigate the convergence of the marginal distributions of the mean field dynamics in the $L^p(m_\infty)$-norm for all $p \in \mathbb{R}$ and take special care when $p < 1$. Let μ be a probability measure on $\mathbb{R}^d$ and $h : \mathbb{R}^d \to [0, +\infty]$ be a measurable function. For $p \neq 0$ define

$$\|h\|_{L^p(\mu)} = \left(\int h(x)^p \mu(\mathrm{d}x)\right)^{1/p},$$

and for $p = 0$ define

$$\|h\|_{L^0(\mu)} = \exp\left(\int \log h(x) \mu(\mathrm{d}x)\right).$$

We say $h \in L^p_+(\mu)$ if

$$\|h\|_{L^p(\mu)} \begin{cases} < +\infty & \text{if } p > 0, \\ \in (0, +\infty) & \text{if } p = 0, \\ > 0 & \text{if } p < 0. \end{cases}$$

It is well-known that $p \mapsto \|h\|_p$ is increasing, which ensures that the 0-norm is well defined once $\|h\|_p < +\infty$ for some $p > 0$ or $\|h\|_q > 0$ for some $q < 0$. In this paper we will only work with μ equal to m_∞, the mean field invariant measure. In this case we write $\|h\|_p = \|h\|_{L^p(m_\infty)}$ for simplicity. We also say $h \in L^{1+}(m_\infty)$ or h is L^{1+}-integrable if there exists a number $p_0 > 1$ such that $h \in L^{p_0}(m_\infty)$.

Statement of main results. Recall that m_t and m^N_t are the respective marginal distributions of the mean field and the N-particle system (1.10), (1.11). We slightly improve the exponential energy dissipation result for the MFL dynamics (1.10).

Theorem 1.4 (Energy dissipation of MFL)**.** *Assume F satisfies* (1.2), (1.3), (1.4) *and* (1.5)*. If m_{t_0} has finite entropy and finite second moment for some $t_0 \geqslant 0$, then for every $t \geqslant t_0$,*

$$H(m_t|m_\infty) \leqslant \mathcal{F}(m_t) - \mathcal{F}(m_\infty) \leqslant \bigl(\mathcal{F}(m_{t_0}) - \mathcal{F}(m_\infty)\bigr)e^{-4\rho(t-t_0)}. \tag{1.18}$$

Remark 1.5. The theorem stated here differs slightly from the previous results ([56, Theorem 3.2] and [178, Theorem 1]), in that we have removed the technical restriction that $x \mapsto D_m F(m, x)$ is infinitely differentiable. This is achieved by using the differential calculus in the Wasserstein space developed in the monograph [4].

The proof of the theorem is postponed to Section 1.4.2.

We also study the MFL system's convergence beyond the entropic sense. In particular, we show that the system converges in the L^2 sense given L^2-initial values (Proposition 1.6), and that the system is also hypercontractive and reverse-hypercontractive (Proposition 1.7).

Denote

$$h_t(x) := \frac{\mathrm{d}m_t}{\mathrm{d}m_\infty}(x)$$

for the solution m_t of the MFL dynamics (1.12), where m_∞ is the unique invariant measure to the MFL, satisfying (1.15).

Proposition 1.6 (L^2-convergence)**.** *Assume F satisfies (1.2), (1.3), (1.4), (1.5) and (1.6). Let $m_t \in C\big([0,+\infty);(\mathcal{P}_2,W_2)\big)$ be a solution to (1.12). If $h_{t_0} \in L^2(m_\infty)$, then $h_t \in L^2(m_\infty)$ for all $t \geqslant t_0$. Moreover, for all $\rho' \in (0,\rho)$, we have*

$$\forall t \geqslant t_0, \qquad \|h_t - 1\|_2^2 \leqslant M e^{-4\rho'(t-t_0)}, \tag{1.19}$$

for the constant M defined by

$$M = \exp\bigg(\frac{\Delta(t_0)}{4\rho}\bigg)\bigg(\|h_{t_0} - 1\|_2^2 + \frac{\Delta(t_0)}{4(\rho-\rho')}\bigg),$$

where

$$\Delta(t_0) = \frac{(M_{mm}^F)^2}{\rho-\rho'}\bigg(1 + \frac{M_{mm}^F}{\rho} + \frac{(M_{mm}^F)^2}{2\rho^2}\bigg)\log\|h_{t_0}\|_2.$$

Proposition 1.7 (Hypercontractivity)**.** *Assume F satisfies (1.2), (1.3), (1.4), (1.5) and (1.6). Suppose $h_{t_0} \in L^{q_0}(m_\infty)$ for some $q_0 \neq 1$. Let $\varepsilon \in (0,1]$ and $q(t)$ solve the ODE $\dot q = 4(1-\varepsilon)\rho(q-1)$ with the initial condition $q(t_0) = q_0$. Then $h_t \in L^{q(t)}(m_\infty)$ for $t \geqslant t_0$. Moreover, we have for $q_0 > 1$,*

$$\log\|h_t\|_{q(t)} \leqslant \log\|h_{t_0}\|_{q_0} + \int_{t_0}^t \delta(s)\,\mathrm{d}s, \tag{1.20}$$

and for $q_0 < 1$,

$$\log\|h_t\|_{q(t)} \geqslant \log\|h_{t_0}\|_{q_0} + \int_{t_0}^t \delta(s)\,\mathrm{d}s, \tag{1.21}$$

where $\delta(t) = \frac{1}{4\varepsilon}\big(q(t)-1\big)(M_{mm}^F)^2 W_1^2(m_t,m_\infty)$.

Remark 1.8 (Optimality of exponent's growth). In the case where the mean field interaction is absent, Nelson's theorem [6, Théorème 2.3.1] shows the optimality of the exponent's growth in Proposition 1.7.

The proofs of Propositions 1.6 and 1.7 are given in Section 1.4.3.

By combining the L^2-convergence and the hypercontractivity, we can obtain the L^p-convergence of the MFL dynamics.

Theorem 1.9 (L^p-convergence of MFL)**.** *Assume F satisfies (1.2), (1.3), (1.4), (1.5) and (1.6). Suppose $h_0 \in L^{p_0}(m_\infty)$ for some $p_0 > 1$. For $\rho' \in (0,\rho)$ and $p \in \mathbb{R}$, we set*

$$\tau_p = \begin{cases} \frac{1}{4\rho'}\log\frac{(p-1)\vee 1}{(p_0-1)\wedge 1}, & \text{if } p \geqslant 0,\\ \frac{1}{4\rho'}\log\frac{2(1-p)}{(p_0-1)\wedge 1}, & \text{if } p < 0.\end{cases}$$

Then for all $t \geqslant \tau_p$, we have that h_t belongs to $L^p(m_\infty)$ and its norm satisfies

$$\begin{aligned}\bigl|\log\|h_t\|_p\bigr| \leqslant \Bigl(\frac{2(1-p)}{p}\mathbb{1}_{p\in(0,1)} + \mathbb{1}_{p\notin(0,1)}\Bigr)&\\ \Bigl(1+\frac{P(\alpha)}{8\varepsilon^2}\Bigr)H_1^{P(\alpha)/4\varepsilon}\bigl(H_1^2-1\bigr)e^{-4(1-\varepsilon)\rho(t-\tau_p)}&\\ + \bigl((p-2)_+\mathbb{1}_{p>0} + (1/2-p)\mathbb{1}_{p\leqslant 0}\bigr)\frac{p_0P(\alpha)\log\|h_0\|_{p_0}}{16(p_0-1)\varepsilon(1-\varepsilon)},&\end{aligned}\tag{1.22}$$

where $\alpha = M_{mm}^F/\rho$, $P(\alpha) = \alpha^2+\alpha^3+\alpha^4/2$, and

$$\log H_1 = \Bigl(1+\frac{p_0(2-p_0)_+P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\Bigr)\log\|h_0\|_{p_0}.$$

Remark 1.10 (Necessity of L^{1+}-initial condition). We here explain why it is necessary to assume $m_0 \in L^{p_0}(m_\infty)$ for some $p_0 > 1$ in Theorem 1.9. Let $m_0(\mathrm{d}x) \propto \exp\bigl(-\sum_{\nu=1}^d |x^\nu|\bigr)\,\mathrm{d}x$, i.e., the d-tensorized exponential distribution and $F(m) = \frac{1}{2}\int |x|^2 m(\mathrm{d}x)$. The Langevin dynamics (1.10) is nothing but Ornstein–Uhlenbeck:

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

The SDE is solved explicitly by

$$X_t = e^{-t}X_0 + \sqrt{2}\int_0^t e^{-(t-s)}\,\mathrm{d}W_s \overset{d}{=} e^{-t}X_0 + \sqrt{1-e^{-2t}}\mathcal{N},$$

where $\mathcal{N} \sim \mathcal{N}(0,1)$ is a standard normal independent from X_0. The Langevin has unique invariant measure $m_\infty \propto \exp(-|x|^2/2)$, i.e., the standard normal distribution in $\mathbb{R}^d$. The initial condition m_0 lies in all $\mathcal{P}_p$ for all $p \geqslant 1$ but m_0/m_∞ does not belong to L^{p_0} for any $p_0 > 1$. And so is m_t. Indeed, for all $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}[\exp(\varepsilon|X_t|^2)] &= \mathbb{E}\bigl[\exp\bigl(\varepsilon(e^{-t}|X_0| + \sqrt{1-e^{-2t}}\mathcal{N})^2\bigr)\bigr]\\ &\geqslant \mathbb{E}\Bigl[\exp\Bigl(\frac{\varepsilon}{2}(e^{-2t}|X_0|^2 - 2(1-e^{-2t})\mathcal{N}^2)\Bigr)\Bigr]\\ &= \mathbb{E}\Bigl[\exp\Bigl(\frac{\varepsilon}{2}e^{-2t}|X_0|^2\Bigr)\Bigr]\,\mathbb{E}\bigl[\exp\bigl(-\varepsilon(1-e^{-2t})\mathcal{N}^2\bigr)\bigr] = +\infty.\end{aligned}$$

Here we used $(a+b)^2 \geqslant \frac{1}{2}a^2 - b^2$ and the independence between X_0 and $\mathcal{N}$. This implies $\int m_t(x)m_\infty(x)^{-\varepsilon}\,\mathrm{d}x = +\infty$ for all $\varepsilon > 0$. Let $p > 1$. By Hölder's inequality we have

$$\begin{aligned}\Bigl(\int m_t(x)^p m_\infty(x)^{-(p-1)}\,\mathrm{d}x\Bigr)^{1/p}\Bigl(\int m_\infty(x)^{1-\varepsilon}\,\mathrm{d}x\Bigr)^{1-1/p}&\\ \geqslant \int m_t(x)m_\infty(x)^{-\varepsilon(1-1/p)}\,\mathrm{d}x = +\infty.&\end{aligned}$$

Hence $\int m_t(x)^p m_\infty(x)^{-(p-1)}\,\mathrm{d}x = +\infty$.

As a by-product of our L^p-convergence result above, we can use the transport method to show the following uniform-in-time concentration of measure result.

Theorem 1.11 (Uniform-in-time concentration of measure). *Under the hypotheses of Theorem 1.9, for all $\rho' \in (0, \rho)$ there exist constants*

$$C_{\rho'} = C_{\rho'}(\rho, M^F_{mm}, p_0, \|h_0\|_{p_0}), \quad \tau_{\rho'} = \tau_{\rho'}(\rho, p_0)$$

such that for every 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, every $t \geqslant \tau_{\rho'}$ and every $r \geqslant 0$,

$$m_t[|f - \mathbb{E}_{m_t} f| \geqslant r] \leqslant 2\exp\Big(-\rho' r^2 + C_{\rho'} e^{-4\rho' t}(r+1)\Big). \tag{1.23}$$

The proofs of Theorems 1.9 1.11 are postponed to Section 1.4.4.

We further study the system of N particles, and show that its marginal distributions approximate $m_\infty^{\otimes N}$, the N-tensorized mean field invariant measure, at a uniform-in-N exponential rate with a uniform-in-N "bias", whose precise meaning will be given below.

Theorem 1.12 (Uniform-in-N energy dissipation of particle systems). *Assume F satisfies (1.2), (1.3), (1.4) and (1.5). If $m^N_{t_0}$ belongs to $\mathcal{P}_2(\mathbb{R}^{dN})$ and has finite entropy for some $N \geqslant 2$ and $t_0 \geqslant 0$, then for all $\rho' \in (0, \rho)$, we have*

$$\begin{aligned} H(m^N_t | m_\infty^{\otimes N}) &\leqslant \mathcal{F}^N(m^N_t) - N\mathcal{F}(m_\infty) \\ &\leqslant \big(\mathcal{F}^N(m^N_{t_0}) - N\mathcal{F}(m_\infty)\big)e^{-(4\rho' - C_1 N^{-1})(t-t_0)} \\ &\qquad + \frac{C_2}{4\rho' - C_1 N^{-1}}, \end{aligned} \tag{1.24}$$

for every $t \geqslant t_0$ and every $N > C_1/4\rho'$, where the constants C_1, C_2 are defined by

$$\begin{aligned} C_1 &= M^F_{mm}\bigg(16 + \frac{6M^F_{mm}\rho'}{\rho(\rho-\rho')}\bigg), \\ C_2 &= dM^F_{mm}\bigg(10 + \frac{3M^F_{mm}\rho'}{\rho(\rho-\rho')}\bigg). \end{aligned}$$

The proof of Theorem 1.12 is postponed to Section 1.5.1.

Remark 1.13 (Sharpness of the size of bias). Let the initial condition m^N_0 of the N-particle system be equal to m^N_∞, the system's invariant measure. By sending t to infinity in (1.24), we have

$$H(m^N_\infty | m_\infty^{\otimes N}) \leqslant \frac{C_2}{4\rho' - C_1 N^{-1}},$$

provided that $\mathcal{F}^N(m^N_\infty) < +\infty$ and $N > C_1/4\rho'$. Drawing an analogy with statistics, we will refer to the relative entropy $H(m^N_\infty | m_\infty^{\otimes N})$ as the 'bias'. Then, the $O(1)$ order of the bias when $N \to +\infty$ is sharp and we give an example attaining this order in the following. Consider the mean field functional

$$F(m) = \frac{1}{2}\int x^2 m(\mathrm{d}x) + \frac{\alpha}{2}\bigg(\int x m(\mathrm{d}x)\bigg)^2$$

with $\alpha \geqslant 0$. We can easily verify all our assumptions on F. The mean field invariant measure is nothing but the d-dimensional standard Gaussian variable:

$$m_\infty(\mathrm{d}x) = (2\pi)^{-d/2}\exp\bigg(-\frac{|x|^2}{2}\bigg)\,\mathrm{d}x,$$

and the invariant measure of the N-particle system reads

$$m_\infty^N(\mathrm{d}\boldsymbol{x}) = (2\pi)^{-dN/2}(\det A_N)^{1/2}\exp\left(-\frac{1}{2}\sum_{i=1}^N |x^i|^2 - \frac{\alpha}{2N}\left(\sum_{j=1}^N x^i\right)^2\right)\mathrm{d}\boldsymbol{x},$$

where A_N is the $Nd \times Nd$ matrix whose $d \times d$ blocks read

$$(A_N)_{ij} = \begin{cases} \left(1+\frac{\alpha}{N}\right)\mathbb{1}_{d\times d} & \text{if } i = j, \\ \frac{\alpha}{N}\mathbb{1}_{d\times d} & \text{if } i \neq j. \end{cases}$$

Especially, we have $\mathcal{F}^N(m_\infty^N) < +\infty$. By diagonalizing A_N, we can obtain $\det A_N = (1+\alpha)^d$. Hence, the relative density between m_∞^N and $m_\infty^{\otimes N}$ reads

$$\frac{\mathrm{d}m_\infty^N}{\mathrm{d}m_\infty^{\otimes N}}(\boldsymbol{x}) = (1+\alpha)^{d/2}\exp\left(-\frac{\alpha}{2N}\left(\sum_{j=1}^N x^i\right)^2\right),$$

and the relative entropy satisfies

$$\begin{aligned} H(m_\infty^N|m_\infty^{\otimes N}) &= \mathbb{E}^{\boldsymbol{X}\sim m_\infty^N}\left[\log\frac{dm_\infty^N}{dm_\infty^{\otimes N}}(\boldsymbol{X})\right] \\ &= \frac{d}{2}\log(1+\alpha) - \frac{\alpha}{2N}\,\mathbb{E}^{\boldsymbol{X}\sim m_\infty^N}\left[\left(\sum_{i=1}^N X^i\right)^2\right] \\ &= \frac{d}{2}\log(1+\alpha) - \frac{d\alpha}{2(1+\alpha)}. \end{aligned}$$

So the $O(1)$ order in N of the bias in (1.24) is sharp.

Finally, we study the propagation of chaos phenomenon. On finite horizon we use the classical arguments of synchronous coupling and Girsanov's theorem to show that the distance between the particle system m_t^N and the tensorized mean field system $m_t^{\otimes N}$ grows at most exponentially, in the sense of Wasserstein distance and relative entropy. On the other hand, for large time, we control the distance using the long time behavior proved in Theorems 1.4, 1.9 and 1.12.

Theorem 1.14 (Wasserstein and entropic propagation of chaos)**.** *Assume F satisfies (1.2), (1.3), (1.4) and (1.5). Suppose m_0 belongs to $\mathcal{P}_2(\mathbb{R}^d)$, m_0^N belongs to $\mathcal{P}_2(\mathbb{R}^{dN})$ and they both have finite entropy for some $N \geqslant 2$.*

- *Then for all $\rho' \in (0,\rho)$, we have*

$$\begin{aligned} \rho W_2^2(m_t^N, m_t^{\otimes N}) &\leqslant 2N\big(\mathcal{F}(m_0) - \mathcal{F}(m_\infty)\big)e^{-4\rho t} \\ &\quad + 2\big(\mathcal{F}^N(m_0^N) - N\mathcal{F}(m_\infty)\big)e^{-(4\rho' - C_1N^{-1})t} + \frac{2C_2}{4\rho' - C_1N^{-1}}, \end{aligned} \tag{1.25}$$

  for every $t \geqslant 0$ and every $N > C_1/4\rho'$, where the constants C_1, C_2 are the same as in Theorem 1.12. If additionally $m_0 \in \mathcal{P}_6(\mathbb{R}^d)$, then we have

$$\begin{aligned} W_2^2(m_t^N, m_t^{\otimes N}) &\leqslant e^{C_4 t}W_2^2(m_0^N, m_0^{\otimes N}) \\ &\quad + NC_5(e^{C_4 t} - 1)\big(v_6(m_0)^{1/3} + 1\big)\delta_d(N), \end{aligned} \tag{1.26}$$

for every $t \geqslant 0$, where $C_4 = \max\bigl(1 + 3(M^F_{mx})^2 + 3(M^F_{mm})^2, 2M^F_{mx} + 4d/3 + 16/3\bigr)$, and C_5 is a constant depending only on M^F_{mx}, M^F_{mm} and d, the term $v_6(m_0)$ is defined by $v_6(m_0) \coloneqq \int\bigl|x - \int x' m_0(\mathrm{d}x')\bigr|^6 m_0(\mathrm{d}x)$ and the term $\delta_d(N)$ is defined by

$$\delta_d(N) \coloneqq \begin{cases} N^{-1/2} & \text{if } d < 4, \\ N^{-1/2}\log(1+N) & \text{if } d = 4, \\ N^{-2/d} & \text{if } d > 4. \end{cases}$$

- *If additionally (1.6) holds and $h_0 \in L^{p_0}(m_\infty)$ for some $p_0 > 1$, then we have*

$$H(m^N_t | m^{\otimes N}_t) \leqslant NC_3 e^{-4\rho' t} + 2\bigl(\mathcal{F}^N(m^N_0) - N\mathcal{F}(m_\infty)\bigr)e^{-(4\rho' - C_1 N^{-1})t} + \frac{2C_2}{4\rho' - C_1 N^{-1}}, \tag{1.27}$$

for every $t \geqslant \tau$ and every $N > C_1/4\rho'$, for some constants C_3, $\tau \geqslant 0$ depending only on ρ, ρ', M^F_{mm}, p_0 and $\|h_0\|_{L^{p_0}(m_\infty)}$. If additionally $m_0 \in \mathcal{P}_6(\mathbb{R}^d)$ and $H(m^N_0 | m^{\otimes N}_0)$ is both finite, then we have

$$H(m^N_t | m^{\otimes N}_t) \leqslant H(m^N_0 | m^{\otimes N}_0) + NC_5(e^{C_4 t} - 1)\bigl(v_6(m_0)^{1/3} + 1\bigr)\delta_d(N), \tag{1.28}$$

for every $t \geqslant 0$, for possibly different constants C_4, $C_5 > 0$ depending on M^F_{mx}, M^F_{mm} and d.

If the initial error is zero, i.e., $m^N_0 = m^{\otimes N}_0$, we obtain the following result by combining the finite-time and long-time estimates, as in the proof of Corollary 5 of [101].

Corollary 1.15. *Assume F satisfies (1.2), (1.3), (1.4) and (1.5). Suppose $m_0 \in \mathcal{P}_6(\mathbb{R}^d)$, m_0 has finite entropy, and $m^N_0 = m^{\otimes N}_0$. Then there exist constants C, $N_0 > 0$, depending on ρ, M^F_{mm}, M^F_{mx}, m_0 and d, such that for all $N \geqslant N_0$,*

$$\sup_{t \in [0,\infty)} \frac{1}{N} W_2^2(m^N_t, m^{\otimes N}_t) \leqslant \frac{C}{N^\kappa} \tag{1.29}$$

where $\kappa = \min(2\rho/C_4, 1)/(d \vee 4)$ with C_4 being the constant in the Wasserstein case of Theorem 1.14. If additionally F satisfies (1.6), we have as well

$$\sup_{t \in [0,\infty)} \frac{1}{N} H(m^N_t | m^{\otimes N}_t) \leqslant \frac{C}{N^\kappa} \tag{1.30}$$

for every $N \geqslant N_0$, with the constants C, κ, $N_0 > 0$ redefined accordingly.

The proofs of Theorem 1.14 and Corollay 1.15 are postponed to Section 1.5.2. The rate κ obtained in the corollary above seems to be highly optimal compared to the $O(1/N)$ rate in Theorem 1.12. This is due to the fact that, for finite time, we do not exploit at all the coercive structure of the MFL. We note that it is recently shown in [70] that if we consider a weaker distance and work under stronger regularity conditions, then the optimal $O(1/N)$ rate can be achieved even when the supremum over all time is taken.

Comments on the assumptions. The conditions (1.3), (1.5) ensure that the drift is jointly Lipschitz continuous in measure and space, which guarantees the well-posedness of the mean field and the particle system dynamics (1.10), (1.11). This also implies that the flow is AC^2 in L^2-Wasserstein space (refer to Definition 1.36), which coincides with the type of curves studied in [4, Chapter 8]. In particular, the "chain rule" holds true, which yields immediately the energy dissipation (1.48) and (1.58).

The assumptions (1.2), (1.4), which have already appeared in the previous works [56, 178], are key to the exponential convergence of relative entropy of the MFL. They are also used in this work, along with (1.3), to show the exponential entropic convergence of the particle system in Theorem 1.12.

The condition (1.6) is technical in that it does not contribute to any constants in our results. This condition allows us to obtain a simple "standard algebra" of the time-dependent semigroup induced by the MFL and to justify easily the computations in L^p spaces needed to prove Theorem 1.9, which is then used to show Theorem 1.14 and Corollary 1.15. It is possible that our results can also be obtained without the higher-order bounds (for example, by an approximation argument). We, however, choose to work in this setting to avoid excessive technicalities.

1.3 Applications

1.3.1 Sufficient conditions for functional convexity

We propose two criteria for the convexity of mean field functionals. The first criterion treats translationally invariant two-body interactions, i.e., energy functionals of the form:

$$F_{\mathrm{Int}}(m) = \frac{1}{2}\iint V(x-y)m(\mathrm{d}x)m(\mathrm{d}y). \tag{1.31}$$

We have the following modified version of Bochner's theorem.

Theorem 1.16 (Bochner)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be a bounded, continuous and even function. Then, the following conditions are equivalent:*

(i) The functional F_{Int}, defined by (1.31), is convex on $\mathcal{P}(\mathbb{R}^d)$.

(ii) For all signed measure μ on $\mathbb{R}^d$ with zero net mass, i.e., $\int \mathrm{d}\mu = 0$, we have $\iint V(x-y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) \geqslant 0$.

(iii) The Fourier transform $\hat{V}$ of V is the sum of a finite and positive measure on $\mathbb{R}^d \setminus \{0\}$ and a scalar multiple of the Dirac mass δ_0 at zero.

The proof of this modified version of Bochner's theorem is postponed to Appendix A.2.

Example 1.17 (Regularized Coulomb). It is well-known that in dimension $d \geqslant 3$ the Coulomb potential $V_{\mathrm{C}}(x) = 1/\big(d(d-2)c_d|x|^{d-2}\big)$ is the fundamental solution to Laplace's equation, that is to say,

$$\Delta V_{\mathrm{C}} = -\delta_0. \tag{1.32}$$

Hence its Fourier transform $\hat{V}_{\mathrm{C}}$ verifies $\hat{V}_{\mathrm{C}}(k) = (2\pi)^{-d/2}|k|^{-2} \geqslant 0$. However $\hat{V}_{\mathrm{C}} \notin L^1(\mathbb{R}^d)$ and Theorem 1.16 does not apply (which is consistent with the singularity of V_{C} at 0). To solve this problem, we propose the regularization

$$\hat{V}_{\mathrm{RC}}(k) = \frac{e^{-r_0|k|}}{(2\pi)^{d/2}|k|^2}$$

for some $r_0 > 0$. Its Fourier inverse $V_{\mathrm{RC}} : \mathbb{R}^d \to \mathbb{R}$ is then indeed a bounded continuous function and has the explicit expression for $d = 3$:

$$V_{\mathrm{RC}}(x) = \int \frac{e^{-r_0|k|}e^{ik\cdot x}}{(2\pi)^3|k|^2}\,\mathrm{d}^3k = \begin{cases} \arctan(|x|/r_0)(2\pi^2|x|)^{-1} & \text{if } x \neq 0, \\ (2\pi^2 r_0)^{-1} & \text{if } x = 0. \end{cases}$$

Note that when $r_0 \to 0$, we have $V_{\mathrm{RC}}(x) \to V_{\mathrm{C}}(x)$ for every $x \in \mathbb{R}^d$. The functional

$$\begin{aligned} F_{\mathrm{RC}}(m) &= \frac{1}{2}\iint V_{\mathrm{RC}}(x-y)m(\mathrm{d}x)m(\mathrm{d}y) \\ &= \frac{1}{2}\iint \frac{1}{2\pi^2}\frac{\arctan(|x-y|/r_0)}{|x-y|}m(\mathrm{d}x)m(\mathrm{d}y) \end{aligned} \tag{1.33}$$

is well defined and convex on $\mathcal{P}(\mathbb{R}^3)$ by Theorem 1.16.

Remark 1.18 (Exclusion of two notions of convexity). If the functional F_{Int} satisfies the conditions of Theorem 1.16, we know

$$2V(0) - V(s) - V(-s) = \frac{2}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\big(1-\cos(k\cdot s)\big)\hat{V}(\mathrm{d}k) \geqslant 0.$$

If the function V is not constant, then there exists some $s_0 \in \mathbb{R}^d$ such that $V(s_0) \neq V(0)$. The evenness of V implies $V(-s_0) = V(s_0)$ and, therefore, $V(s_0) = V(-s_0) < V(0)$. In particular, V is not convex, and the functional F_{Int} cannot be geodesically convex. In other words, the only functionals of form (1.31) with continuous, bounded and even V that are both functionally and geodesically convex are constant functionals.

Remark 1.19. Other regularizations preserving the positivity of the Coulomb potential can also be possible. For example we can convolute Laplace's equation (1.32) with a heat kernel $\rho^\varepsilon : x \mapsto (2\pi\varepsilon)^{-d/2}\exp\big(-(2\varepsilon)^{-1}x^2\big)$ to obtain

$$\Delta V'_{\mathrm{RC}} = \Delta(V_{\mathrm{C}} \star \rho^\varepsilon) = -\rho^\varepsilon.$$

The Fourier transform of V'_{RC} reads

$$\hat{V}'_{\mathrm{RC}}(k) = \frac{\hat{\rho}^\varepsilon(k)}{|k|^2} = \frac{e^{-2\pi^2\varepsilon|k|^2}}{(2\pi)^{d/2}|k|^2},$$

which is positive and L^1-integrable. The main reason for choosing the regularization in Example 1.17 is that it allows for the simple expression given in (1.33) in three dimensions.

The second criterion is an analogue of the property of convex functions under composition.

Proposition 1.20. *Let X be a Banach space. If $V : \mathbb{R}^d \to X$ is a function of quadratic growth and $g : X \to \mathbb{R}$ is convex, then the functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined by*

$$F(m) = g\bigg(\int V(x)m(\mathrm{d}x)\bigg)$$

is convex.

Proof. Immediate. □

Example 1.21 (L^2-loss of two-layer neural networks). We first explain the structure of two-layer neural networks and then introduce the mean field model for it. Consider an *activation function* $\varphi : \mathbb{R} \to \mathbb{R}$ satisfying

$$\begin{gathered}\varphi \text{ is continuous and non-decreasing,} \\ \lim_{x\to-\infty} \varphi(x) = 0, \qquad \lim_{x\to+\infty} \varphi(x) = 1,\end{gathered} \tag{1.34}$$

Define $S = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$, where the *neurons* take values. For each neuron $\theta = (c, a, b) \in S$ we define the *feature map*:

$$\mathbb{R}^d \ni z \mapsto \Phi(\theta; z) \coloneqq \ell(c)\varphi(a \cdot z + b) \in \mathbb{R}, \tag{1.35}$$

where $\ell : \mathbb{R} \to [-L, L]$ is a *truncation function* with the *truncation threshold* $L \in (0, +\infty]$. Such truncation has been considered in recent papers [117, 178]. The two-layer neural network is nothing but the averaged feature map parameterized by N neurons $\theta^1, \dots, \theta^N \in S$:

$$\mathbb{R}^d \ni z \mapsto \Phi^N(\theta^1, \dots, \theta^N; z) = \frac{1}{N} \sum_{i=1}^N \Phi(\theta^i; z) \in \mathbb{R}. \tag{1.36}$$

The training of neural network aims to minimize the distance between the averaged output (1.36) and a (only empirically known) *label* function $f : \mathbb{R}^d \to \mathbb{R}$, i.e.

$$\inf_{(\theta^1,\dots,\theta^N)\in S^N} \boldsymbol{d}\big(f, \Phi^N(\theta^1, \dots, \theta^N; \cdot)\big) \tag{1.37}$$

for some loss functional $\boldsymbol{d}$. In this paper, we use the $L^2(\mu)$-norm as the loss functional where $\mu \in \mathcal{P}(\mathbb{R}^d)$ represents the *feature* distribution. In this way, the objective function of the minimization reads

$$F^N_{\mathrm{NNet}}(\theta^1, \dots, \theta^N) = \frac{N}{2} \int \big|f(z) - \Phi^N(\theta^1, \dots, \theta^N; z)\big|^2 \mu(\mathrm{d}z). \tag{1.38}$$

To fit the problem to our theoretical framework, we assume that the feature map $\Phi : S \times \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\begin{aligned} \forall \theta \in S, &\qquad \Phi(\theta; \cdot) \in L^2(\mu), \\ \exists C > 0,\ \forall \theta \in S, &\qquad \|\Phi(\theta; \cdot)\|_{L^2(\mu)} \leqslant C(1 + |\theta|^2). \end{aligned}$$

Now we present the mean field formulation of two-layer neural networks. Let $\mathcal{P}_2(S)$ be the space of probability measures on S of finite second moment and define the class of functions representable by the mean field neural network by:

$$\mathcal{N}_{\varphi,\ell} = \{h : \mathbb{R}^d \to \mathbb{R} : \exists m \in \mathcal{P}_2(S),\ \forall x \in \mathbb{R}^d, \quad h(x) = \mathbb{E}^{\Theta\sim m}[\Phi(\Theta; x)]\}. \tag{1.39}$$

In particular the N-neuron output functions defined in (1.36) belong to this class since

$$\Phi^N(\theta^1, \dots, \theta^N; \cdot) = \mathbb{E}^{\Theta\sim \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i}}[\Phi(\Theta; \cdot)].$$

Instead of the finite-dimensional optimization (1.37), we consider the following mean field optimization:

$$\inf_{\mathcal{P}_2(S)} F_{\mathrm{NNet}}(m),$$
$$\text{where } F_{\mathrm{NNet}}(m) := \boldsymbol{d}\big(f, \mathbb{E}^{\Theta\sim m}[\Phi(\Theta;\cdot)]\big) = \frac{1}{2}\int \big|f(z) - \mathbb{E}^{\Theta\sim m}[\Phi(\Theta;z)]\big|^2 \mu(\mathrm{d}z). \tag{1.40}$$

The functional F_{NNet} is convex by Proposition 1.20 since

$$F_{\mathrm{NNet}}(m) = g\bigg(\int V(\theta)m(\mathrm{d}\theta)\bigg)$$

with $V : S \ni \theta \mapsto \big(z \mapsto \Phi(\theta;z)\big) \in L^2(\mu)$ and $g : L^2(\mu) \ni h \mapsto \|f-h\|^2_{L^2(\mu)} \in \mathbb{R}$.

Remark 1.22 (Motivation of mean field formulation). The N-neuron problem (1.38) is non-convex due to the non-linear activation function φ. Inspired by the fact that the width N of two-layer neural networks is usually large in practice, the authors of [163, 57, 203, 117] consider the mean field formulation of neural networks which convexifies the original problem.

Remark 1.23 (Absence of geodesic convexity). We highlight here that if F_{NNet} is geodesically convex and regular enough, then the N-neuron problem F^N_{NNet} is convex, which is not true. Hence by contradiction F_{NNet} has no geodesic convexity. Indeed, suppose F_{NNet} is geodesically convex. Note that $t \mapsto \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i + tv^i}$ is a geodesic in $(\mathcal{P}_2, W_2)$ in a neighborhood of $t = 0$ if θ_i are distinct from each other (as the pairing $(\theta^i, \theta^i + tv^i)$, $i = 1, \ldots, N$ verifies cyclical monotonicity for t small enough). By the geodesic convexity of F_{NNet} and the relation $F^N_{\mathrm{NNet}}(\theta^1, \ldots, \theta^N) = NF_{\mathrm{NNet}}\big(\frac{1}{N}\sum_{i=1}^N \delta_{\theta^i}\big)$, we obtain the local convexity of F^N_{NNet} on the set

$$S^N \setminus \Delta^N := S^N \setminus \{(\theta_1, \ldots, \theta_N) \in S^N : \exists i \neq j, \quad \theta_i = \theta_j\}.$$

If F^N_{NNet} is additionally C^2, the local convexity implies $\nabla^2 F^N_{\mathrm{NNet}} \geqslant 0$ on $S^N \setminus \Delta^N$ and by density $\nabla^2 F^N_{\mathrm{NNet}} \geqslant 0$ everywhere. Therefore F^N_{NNet} is convex on S^N.

Remark 1.24 (Expressiveness of truncated networks). It is well known that two-layer neural networks are universal approximators, that is, they can approximate any continuous function on $\mathbb{R}^d$ arbitrarily well with respect to the compact-open topology ([116, Theorem 2.4]). This implies that the infimum in (1.40) is zero if μ is compactly supported and no truncation is present (that is, $L = +\infty$ and ℓ is the identity function). However, if a truncation with $L < +\infty$ is applied, all functions $h \in \mathcal{N}_{\varphi,\ell}$ satisfy the bound $\|h\|_\infty \leqslant L$ and therefore cannot approximate well functions that exceed L. However, Barron's theorem [13, Theorem 2] says that if a function f satisfies

$$f(x) = f(0) + \int (e^{i\omega\cdot x} - 1)F(\mathrm{d}\omega)$$

for every $x \in B(0,R)$, for some complex-valued measure F, and if there exists c_+, $c_- \in \mathbb{R}$ such that $\ell(c_+) = L$ and $\ell(c_-) = -L$, and that

$$L \geqslant R\int |\omega||F(\mathrm{d}\omega)| + |f(0)|,$$

then the best approximation error

$$\inf_{\Phi\in\mathcal{N}_{\varphi,\ell}} \|f-\Phi\|_{L^2(\mu)} = 0$$

for every probability measure μ supported in $B(0,R)$.

1.3.2 Examples of MFL dynamics

We construct MFL dynamics for the two examples discussed earlier and demonstrate that our theorems are applicable in both cases. To verify the LSI condition (1.4) we will use the following results.

Proposition 1.25. *Let $\mu(\mathrm{d}x) = e^{-V(x)}\,\mathrm{d}x$ be a probability measure in $\mathbb{R}^d$ for some $V \in C^2(\mathbb{R}^d)$.*

- *(Bakry–Émery [11]) If $\nabla^2 V \geqslant \kappa$ then μ satisfies a $\kappa/2$-LSI.*
- *(Holley–Stroock [113]) If $V = V_1 + V_2$, where e^{-V_1} is the density of a probability measure satisfying an ρ-LSI and V_2 is bounded with oscillation $\operatorname{osc} V_2$, then μ satisfies a $\rho\exp(-\operatorname{osc} V_2)$-LSI.*
- *(Aida–Shigekawa [1]) If V_2 in the previous statement is Lipschitz-continuous instead of bounded, then μ satisfies an LSI as well.*

Example 1.26 (MFL for regularized Coulomb system). Let $\lambda > 0$. Define

$$F_{\mathrm{Ext}}(m) = \frac{\lambda}{2}\int |x|^2 m(\mathrm{d}x). \tag{1.41}$$

We consider the functional $F = F_{\mathrm{RC}} + F_{\mathrm{Ext}}$ where F_{RC} is defined in (1.33). By the discussions in Example 1.17 the functional F satisfies the convexity condition (1.2). Its linear functional derivative reads

$$\frac{\delta F}{\delta m}(m,x) = \int V_{\mathrm{RC}}(x-y)m(\mathrm{d}y) + \frac{1}{2}\lambda|x|^2$$

and its intrinsic derivative reads $D_m F(m,x) = \int \nabla V_{\mathrm{RC}}(x-y)m(\mathrm{d}y) + \lambda x$. The conditions (1.3), (1.5), (1.6) are satisfied because

$$\|\nabla^n V_{\mathrm{RC}}\|_\infty \leqslant \frac{1}{(2\pi)^{d/2}}\int |k|^n \hat{V}_{\mathrm{RC}}(\mathrm{d}k) = \int |k|^n \frac{e^{-r_0|k|}}{(2\pi)^d|k|^2}\,\mathrm{d}^d k < +\infty$$

for all $n \geqslant 0$ (and $d \geqslant 3$). In particular, the bound in (1.3) is verified by $M^F_{mm} = \|\nabla^2 V_{\mathrm{RC}}\|_\infty$. For the uniform LSI, we can apply Holley–Stroock or Aida–Shigakawa, since the first term in $\frac{\delta F}{\delta m}$ is uniformly bounded and uniformly Lipschitz and the second term verifies the Bakry–Émery condition. The LSI constant given by Holley–Stroock has the simple expression in three dimensions $\rho = \lambda\exp(-\operatorname{osc} V_{\mathrm{RC}})/2 = \lambda\exp(-1/2\pi^2 r_0)/2$. The L^{1+}-integrability of the initial condition, needed by Theorem 1.9 and the second part of Theorem 1.14, is verified once we have

$$\exists C,\varepsilon > 0,\ \forall x \in \mathbb{R}, \qquad m_0(x) \leqslant Ce^{-\varepsilon|x|^2}. \tag{1.42}$$

However, as the regularization parameter r_0 approaches 0, we observe $\rho \to 0$ and $M^F_{mm} \to +\infty$, suggesting our method is not suitable for the unregularized Coulomb interaction. We refer readers to [32, 33, 199, 59] for recent developments on the noised gradient flow of Coulomb (and more generally, Riesz) particle systems, where the *modulated free energy* is used to tackle the singularity in the interactions.

Example 1.27 (MFL for two-layer neural networks). Recall the mean field two-layer neural networks in Example 1.21. Suppose

- the truncation L is finite;
- the activation and truncation functions φ, ℓ have bounded derivatives of up to fourth order;
- the feature distribution μ has finite second moment;
- the label function f belongs to $L^2(\mu)$.

On top of the mean field optimization problems (1.40), we add the quadratic regularizer F_{Ext} in (1.41) to the loss, as for the Coulomb system. Then the function and the functional to optimize read

$$\begin{aligned} F^N(\theta^1,\dots,\theta^N) &= \frac{N}{2}\int\biggl|f(z)-\frac{1}{N}\sum_{i=1}^N\Phi(\theta^i;z)\biggr|^2\mu(\mathrm{d}z)+\frac{\lambda}{2}\sum_{i=1}^N|\theta^i|^2,\\ F(m) &= \frac{1}{2}\int\bigl|f(z)-\mathbb{E}^{\Theta\sim m}[\Phi(\Theta;z)]\bigr|^2\mu(\mathrm{d}z)+\frac{\lambda}{2}\int|\theta|^2m(\mathrm{d}\theta). \end{aligned}$$

The N-neuron loss can be recover from the mean field loss by $F^N(\theta^1,\dots,\theta^N)=NF\bigl(\frac{1}{N}\sum_{i=1}^N\delta_{\theta^i}\bigr)$. We verify the assumptions of our theorems one by one. The functional convexity of $F=F_{\mathrm{NNet}}+F_{\mathrm{Ext}}$ is already proved in Example 1.21. The linear functional derivative of F reads

$$\frac{\delta F}{\delta m}(m,\theta)=-\int\bigl(f(z)-\mathbb{E}^{\Theta\sim m}[\Phi(\Theta;z)]\bigr)\Phi(\theta;z)\mu(\mathrm{d}z)+\frac{\lambda}{2}|\theta|^2.$$

The first term on the right hand side is uniformly bounded: for every $m\in\mathcal{P}_2(S)$ and every $\theta\in S$,

$$\biggl|\int\bigl(f(z)-\mathbb{E}^{\Theta\sim m}[\Phi(\Theta;z)]\bigr)\Phi(\theta;z)\mu(\mathrm{d}z)\biggr|\leqslant(\|f\|_{L^1(\mu)}+\|\ell\|_\infty)\|\ell\|_\infty.$$

Hence by Holley–Stroock the uniform LSI condition (1.4) is satisfied with the constant

$$\rho=\frac{\lambda}{2}\exp\bigl(-2(\|f\|_{L^1(\mu)}+\|\ell\|_\infty)\|\ell\|_\infty\bigr).$$

The intrinsic derivative of F reads

$$D_mF(m,\theta)=-\int\bigl(f(z)-\mathbb{E}^{\Theta\sim m}[\Phi(\Theta;z)]\bigr)\frac{\partial\Phi}{\partial\theta}(\theta;z)\mu(\mathrm{d}z)+\lambda\theta,$$

where the partial derivative of the feature map Φ, defined in (1.35), reads

$$\frac{\partial\Phi}{\partial c}(\theta;z)=\ell'(c)\varphi(a\cdot z+b),\ \frac{\partial\Phi}{\partial a}(\theta;z)=\ell(c)\varphi'(a\cdot z+b)z,\ \frac{\partial\Phi}{\partial b}(\theta;z)=\ell(c)\varphi'(a\cdot z+b)$$

for $\theta=(c,a,b)\in S$. Similarly we obtain the second order intrinsic derivative: $D_m^2F(m,\theta,\theta')=\int\frac{\partial\Phi}{\partial\theta}(\theta;z)\otimes\frac{\partial\Phi}{\partial\theta}(\theta';z)\mu(\mathrm{d}z)$. Its 2-norm satisfies the bound

$$|D_m^2F(m,\theta,\theta')|_2^2\leqslant\|\ell'\|_\infty^2+\|\ell\|_\infty^2\|\varphi'\|_\infty^2\bigl(1+M_2(\mu)\bigr),$$

where $M_2(\mu) = \int |z|^2\mu(\mathrm{d}z)$ is the second moment of μ. Thanks to the Kantorovich duality and the Cauchy–Schwarz inequality, the W_1-Lipschitz constant of $m \mapsto D_m F(m,x)$ can be given by

$$M^F_{mm} = \Big(\|\ell'\|^2_\infty + \|\ell\|^2_\infty\|\varphi'\|^2_\infty\big(1+M_2(\mu)\big)\Big)^{1/2}.$$

So D_mF satisfies the condition (1.3). Since ℓ, φ have bounded derivatives of up to fourth order, the derivatives $\nabla^k D_m F(m,\theta)$ for $k = 1, 2, 3$ are also uniformly bounded. Thus the technical conditions (1.5) (1.6) are also satisfied. Finally, the L^{1+}-integrability of the initial value m_0 is verified once we require the pointwise Gaussian bound (1.42) on the density of m_0.

Remark 1.28 (Link to practice). In the training of neural networks, the measure μ is an empirical measure $\frac{1}{K}\sum_{k=1}^K \delta_{z_k}$ and on the feature points $\{z_k\}_{k=1}^K$ the labels are known $f(z_k) = y_k$. This collection of pairs $\{z_k, y_k\}_{k=1}^K$ are the available training data. In practice, instead of the mean field dynamics, we can only simulate the corresponding N-particle system. In other words, we calculate the N-neuron SDE

$$\mathrm{d}\Theta^i_t = \frac{1}{K}\sum_{k=1}^K \big(y_k - \Phi^N(\Theta^1_t,\dots,\Theta^N_t;z_k)\big)\frac{\partial\Phi}{\partial\theta}(\Theta^i_t;z_k)\,\mathrm{d}t - \lambda\Theta^i_t\,\mathrm{d}t + \sigma\,\mathrm{d}W^i_t, \quad (1.43)$$

for $i = 1,\dots,N$. The first drift term of the diffusion is the gradient

$$\nabla_{\theta^i}F^N(\Theta^1_t,\dots,\Theta^N_t),$$

so the time-discretization of this diffusion is nothing but the *noisy gradient descent* (NGD) algorithm for training neural networks. We refer readers to [232, 235, 153, 234, 173] for its applications. The second drift term $-\lambda\Theta^i_t$, coming from our quadratic regularization, is called *weight decay* in the field of machine learning. It is believed to lead to better generalizations of the trained neural network (see [135, 155]).

Remark 1.29 (Noised data). In the previous remark we suppose the data available $\{z_k, y_k\}_{k=1}^N$ are precise: $y_k = f(z_k)$, while in practice they may be subject to errors: $y'_k = f(z_k) + \varepsilon_k$. The new collection of points $\{z_k, y'_k\}_{k=1}^N$ induces another mean field functional F'_{NNet} defined by

$$F'_{\text{NNet}}(m) = \frac{1}{2K}\sum_{k=1}^K \big(y'_k - \mathbb{E}^{\Theta\sim m}[\Phi(\Theta;z_k)]\big)^2.$$

From the triangle inequality for the L^2-distance we deduce

$$|F'_{\text{NNet}}(m) - F_{\text{NNet}}(m)| \leqslant \left(\frac{1}{K}\sum_{k=1}^K \varepsilon_k^2\right)^{1/2} F_{\text{NNet}}(m)^{1/2} + \frac{1}{2K}\sum_{k=1}^K \varepsilon_k^2.$$

The actual N-neuron training process is therefore the noised gradient descent for the functional $F' \coloneqq F'_{\text{NNet}} + F_{\text{Ext}}$ and approximately converges to $(m'_\infty)^{\otimes N}$ where m'_∞ minimizes $\mathcal{F}' = F' + \frac{\sigma^2}{2}H$. The difference between respective minima can be bounded as follows:

$$\begin{aligned}\mathcal{F}'(m'_\infty) - \mathcal{F}(m_\infty) &\leqslant \mathcal{F}'(m_\infty) - \mathcal{F}(m_\infty) = F'_{\text{NNet}}(m_\infty) - F_{\text{NNet}}(m_\infty)\\ &\leqslant \left(\frac{1}{K}\sum_{k=1}^K \varepsilon_k^2\right)^{1/2} F_{\text{NNet}}(m_\infty)^{1/2} + \frac{1}{2K}\sum_{k=1}^K \varepsilon_k^2.\end{aligned}$$

Hence the additional error converges to zero as the noise in the data $(\varepsilon_k)_{k=1}^K$ tends to zero.

Remark 1.30 (Advantages over other approaches). Our Theorems 1.12 and 1.14 establish the exponential convergence of the N-neurons training process (1.43) without supposing the truncation satisfies the regularity conditions such as $\|\nabla^k \ell\|_\infty < c$ for some small constant c. This stands in contrast to many previous studies on uniform-in-time propagation of chaos relying on the smallness of the mean field interaction (e.g. [80] and the first setting of [70]). Yet the smallness approach does not apply to general neural networks: in our setting, the smallness requires the Lipschitz constants M_{mm}^F to be smaller than a constant times ρ, which we denote by $M_{mm}^F \lesssim \rho$, and the relation is difficult to verify. Indeed, using the constants M_{mm}^F, ρ obtained in Example 1.27, we need

$$\Big(\|\ell'\|_\infty^2 + \|\ell\|_\infty^2 \|\varphi'\|_\infty^2 \big(1 + M_2(\mu)\big)\Big)^{1/2} \lesssim \frac{\lambda}{2} \exp\big(-2(\|f\|_{L^1(\mu)} + \|\ell\|_\infty)\|\ell\|_\infty\big).$$

This forces either the regularization λ to be very large or the truncation $\|\ell\|_\infty$ to be very small. In conclusion, our approach based on the functional convexity offers the advantage of obtaining the exponential convergence, albeit at a very slow rate, without such restrictions on λ or ℓ.

1.3.3 Numerical experiments

As explained in Examples 1.21 and 1.27, the MFL dynamics for training two-layer neural networks verifies all the conditions of our theorems, so its particle systems satisfy the uniform exponential energy dissipation (1.24). We now present our numerical experiments.

Setup. We aim to train a neural network to approximate the elementary function $z \mapsto f(z) = \sin 2\pi z_1 + \cos 2\pi z_2$ on $[0,1]^2$. We uniformly sample K points $\{z_i\}_{k=1}^K$ from $[0,1]^2$ and calculate the corresponding labels $y_k = f(z_k)$ to prepare our training data $\{z_k, y_k\}_{k=1}^K$. These points are plotted in Figure 1.1. We fix the truncation function ℓ by $\ell(x) = (x \wedge 100) \vee -100$ and the sigmoid activation function φ by $\varphi(x) = 1/\big(1 + \exp(-x)\big)$. The Brownian noise has volatility σ, and it is necessary to apply the scaling transform in Remark 1.3 before comparing to the theoretical results. Additionally, the quadratic regularization constant λ is fixed in our experiments. The initial values $(\Theta_0^i)_{i=1}^N = (c_0^i, a_0^i, b_0^i)_{i=1}^N$ of the N neurons are sampled independently from a normal distribution m_0 in four dimensions. The training process (1.43) is discretized with time step Δt and terminated at time T. The values of the hyperparameters K, σ, m_0, Δt, T are listed in Table 1.1 and the training algorithm is shown in Algorithm 1. We take the number of neurons N to be 2^P for $P = 6, \dots, 10$ and repeat the training 10 times for each N.

Results. We compute the sum of the N^{-1}-scaled loss $\frac{1}{N} F_{\mathrm{NNet}}^N(\Theta_t^1, \dots, \Theta_t^N)$ at each time t and plot its evolution in Figure 1.2. We observe the value of $\frac{1}{N} F_{\mathrm{NNet}}^N$ first decreases exponentially and then decreases more slowly or even stabilizes. To explore the relationship between this residual error and the number of neurons, for each value of N we calculate the average value of $\frac{1}{N} F_{\mathrm{NNet}}^N$ during the last 500 training steps and take the average of these values over the 10 independent runs. The results are plotted in Figure 1.3.

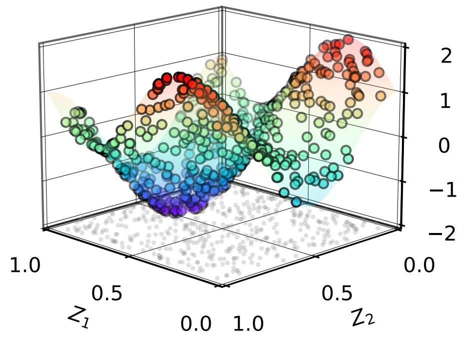


Figure 1.1: Data samples $\{z_k, y_k\}_{k=1}^K$ (schematic).

Parameters	Value
Δt	0.2
T	4000
K	1000
m_0	$\mathcal{N}(0, 5^2)$
σ	1
λ	10^{-5}

Table 1.1: Hyperparameters of neural network training.

Algorithm 1: Noised gradient descent for training a two-layer neural network

Input: number of particles N, activation φ, truncation ℓ, data set $(z_k, y_k)_{k=1}^K$, noise σ, initial distribution m_0, time step Δt, time horizon T

Output: $(\Theta_T^i)_{i=1}^N$

generate i.i.d. $\Theta_0^i = (A_0^i, B_0^i, C_0^i) \sim m_0, i = 1, \dots, N$;

for $t = 0, \Delta t, 2\Delta t, \dots, T - \Delta t$ **do**

  generate i.i.d. $\mathcal{N}_t^i \sim \mathcal{N}(0, 1), i = 1, \dots, N$;

  `// update particles according to discretized Langevin`

  for $i = 1, \dots, N$ **do**

    $\Theta_{t+\Delta t}^i \leftarrow \Theta_t^i - \Big(D_m F_{\text{NNet}}\big(\frac{1}{N}\sum_{j=1}^N \delta_{\Theta_t^j}, \Theta_t^i\big) + \lambda\Theta_t^i\Big)\Delta t + \sigma\sqrt{\Delta t}\,\mathcal{N}_t^i$;

    `/* where` $D_m F_{\text{NNet}}\big(\frac{1}{N}\sum_{j=1}^N \delta_{\Theta_t^j}, \Theta_t^i\big) =$

      $\frac{1}{K}\sum_{k=1}^K \big(y_k - \Phi^N(\Theta_t^1, \dots, \Theta_t^N; z_k)\big)\frac{\partial\Phi}{\partial\theta}(\Theta_t^i; z_k)$ `*/`

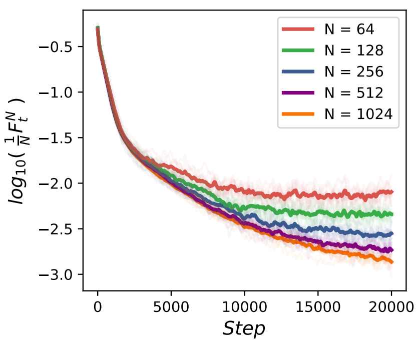


Figure 1.2: Individual (shadowed) and 10-averaged (bold) losses versus time steps.

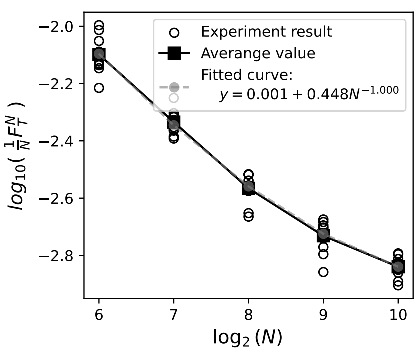


Figure 1.3: Average losses of last 500 steps for individual trainings (shadowed) and its 10-average (bold).

Discussions. Our truncation function ℓ does not have bounded derivatives of up to fourth order as required in Example 1.27 and we can work around this by taking a sequence of regular ℓ_n approximating ℓ since the constants M^F_{mm}, ρ depends only on $\|\ell\|_\infty, \|\ell'\|_\infty$. In our experiment we also ignore the time-discretization error and the difference between training and validation data sets. As shown in Figure 1.2 the losses first decrease exponentially at a uniform rate for different numbers of neurons, N. This is consistent with the convergence rate $\rho' - \frac{C_1}{N}$ predicted by Theorems 1.12 and 1.14. However, the LSI constant obtained in Example 1.27 by Holley–Stroock is excessively small and fails to predict the actual convergence rate. Given that the Holley–Stroock method relies solely on the boundedness of neural networks, this phenomenon suggests the internal structure of neural networks allows for a faster convergence rate that is not captured by the perturbation lemma.

We fit the residual losses with the curve $\frac{\alpha}{N} + \beta$ in Figure 1.3. We choose this parametrization for two reasons: the first term $\frac{\alpha}{N}$ corresponds to the error term in the convergence result (1.24) of the free energy $\frac{1}{N}\mathcal{F}^N(m^N_t)$; the second term β accounts for the facts that $\mathcal{F}(m_\infty) \neq 0$ and that the free energy differs from the neural network's loss by

$$\frac{1}{N}\mathcal{F}^N(m^N_t) - \frac{1}{N}F^N_{\mathrm{NNet}}(m^N_t) = \frac{\lambda}{2N}\int |\boldsymbol{\theta}|^2 m^N_t(\mathrm{d}\boldsymbol{\theta}) + \frac{\sigma^2}{2N}H(m^N_t).$$

In particular the relative entropy $H(m^N_t)$ can not be directly calculated.

1.4 Mean field system

1.4.1 Existence of the measures $\hat{m}$, m_∞, m^N_∞

Our assumptions differ from those in the earlier works, such as [117]. Specifically, we do not require the coercivity condition of type

$$\forall m \in \mathcal{P}_2(\mathbb{R}^d),\ \forall x \in \mathbb{R}^d, \qquad D_m F(m,x)\cdot x \geqslant C(|x|^2 - 1).$$

Instead we only assume the condition (1.5) on $D_mF(m,x)$. As a result, the existence of the measures $\hat{m}, m_\infty, m^N_\infty$, introduced in Section 1.2, is not obvious. In this subsection we show that thanks to the conditions (1.2), (1.3), (1.4), these measures are indeed well defined.

First we sketch a proof that regular enough measures satisfying an LSI in $\mathbb{R}^d$ have finite moments.

Lemma 1.31. *Let $\mu(\mathrm{d}x) = e^{-\Psi}\,\mathrm{d}x$ be a probability measure in $\mathbb{R}^d$ where Ψ is twice differentiable with the bound $|\nabla^2\Psi| \leqslant C$. If μ satisfies an LSI, i.e. (1.4) holds when $\hat{m}$ is replaced by μ for some $\rho > 0$, then $\mu \in \cap_{p\geqslant 1}\mathcal{P}_p(\mathbb{R}^d)$ and $\int e^{\alpha|x|}\mu(\mathrm{d}x) < +\infty$ for all $\alpha \geqslant 0$.*

Proof. Here we repeat the argument of Otto and Villani in [180]. Suppose μ satisfies a ρ-LSI (but we do not suppose $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ a priori). For every measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ of finite entropy (e.g. the Gaussians), the heat flow

$$\partial_t \nu_t = \Delta\nu_t + \nabla\cdot(\nu_t\nabla\Psi), \qquad \nu_0 = \nu$$

is well defined and is an absolutely continuous curve in $(\mathcal{P}_2, W_2)$ thanks to the bound $|\nabla^2\Psi| \leqslant C$ and [22, Theorem 7.4.1]. Hence by the argument of [180, Proposition

1'], we can obtain $H(\nu_t|\mu) \leqslant H(\nu|\mu)e^{-4\rho t}$ and

$$W_2(\nu,\nu_t) \leqslant \frac{1}{\sqrt{\rho}}\Big(\sqrt{H(\nu|\mu)} - \sqrt{H(\nu_t|\mu)}\Big). \tag{1.44}$$

The sequence ν_t are tight in the weak topology of $\mathcal{P}$ since we have $\rho W_2(\nu,\nu_t)^2 \leqslant H(\nu|\mu) = \int(\log\nu + \Psi)\nu < +\infty$ (recall that Ψ is of quadratic growth). By the lower-semicontinuity of $H(\cdot|\mu)$ we must have $\nu_t \to \mu$ in $\mathcal{P}$ weakly when $t \to \infty$. Then we take $\liminf_{t\to\infty}$ on both side of (1.44) and use the lower-semicontinuity of W_2 with respect to the weak topology of $\mathcal{P}$ to obtain Talagrand's inequality

$$\rho W_2^2(\nu,\mu) \leqslant H(\nu|\mu).$$

Hence $\mu \in \mathcal{P}_2$. Finiteness of higher moments and exponential moments then follows from concentration inequalities via Herbst's argument (see e.g. the proof of [27, Theorem 5.5]). $\square$

We give a sufficient condition to the existence of $\hat{m}$ for every $m \in \mathcal{P}_2(\mathbb{R}^d)$ so that the condition (1.4) makes sense.

Proposition 1.32. *Assume F satisfies (1.3). If there exists a measure $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\hat{m}_0$ is well defined (i.e. $Z(\hat{m}_0) < +\infty$) and m_0 satisfies LSI (1.4), then $\hat{m}$ are well defined (i.e. $Z(\hat{m}) < +\infty$) for all $m \in \mathcal{P}_2(\mathbb{R}^d)$.*

Proof. By definition we have

$$\begin{aligned} Z(\hat{m}) &= \int \exp\Big(-\frac{\delta F}{\delta m}(m,x)\Big)\,\mathrm{d}x \\ &= Z(\hat{m}_0)\int \exp\Big(\frac{\delta F}{\delta m}(m_0,x) - \frac{\delta F}{\delta m}(m,x)\Big)\hat{m}_0(\mathrm{d}x), \end{aligned}$$

where the term on the exponential is of linear growth since its derivative is uniformly bounded: $\big|\nabla\big(\frac{\delta F}{\delta m}(m_0,x) - \frac{\delta F}{\delta m}(m,x)\big)\big| = |D_mF(m_0,x) - D_mF(m,x)| \leqslant M_{mm}^F W_2(m_0,m)$. But by Lemma 1.31, all exponential moments of $\hat{m}_0$ are finite. Thus $Z(\hat{m}) < +\infty$ and $\hat{m}$ is well defined. $\square$

We now show that the N-particle invariant measure is also well defined.

Proposition 1.33. *Assume F satisfies (1.2) and (1.4). Then the measure m_∞^N in (1.17) is well defined and has finite exponential moments for all $N \geqslant 2$.*

Proof. Fix $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Using convexity we obtain

$$\begin{aligned} NF(\mu_{\boldsymbol{x}}) &\geqslant NF(m_0) + N\int \frac{\delta F}{\delta m}(m_0,y)(\mu_{\boldsymbol{x}} - m_0)(\mathrm{d}y) \\ &= NF(m_0) - N\int \frac{\delta F}{\delta m}(m_0,y)m_0(\mathrm{d}y) + \sum_{i=1}^N \frac{\delta F}{\delta m}(m_0,x^i). \end{aligned}$$

The integral $\int \frac{\delta F}{\delta m}(m_0,y)m_0(\mathrm{d}y)$ is finite thanks to Lemma 1.31. Hence

$$\int \exp\big(-NF(\mu_{\boldsymbol{x}})\big)\,\mathrm{d}\boldsymbol{x} \leqslant C\int \exp\Big(-\sum_{i=1}^N \frac{\delta F}{\delta m}(m_0,x^i)\Big)\,\mathrm{d}\boldsymbol{x} = C\big(Z(\hat{m}_0)\big)^N < +\infty.$$

Apply the same argument to $\int \exp\big(\alpha\sum_{i=1}^N |x^i|\big)\exp\big(-NF(\mu_{\boldsymbol{x}})\big)\,\mathrm{d}\boldsymbol{x}$ we obtain the finiteness of exponential moments. $\square$

Proposition 1.34. *Assume F satisfies (1.2), (1.4), (1.3) and (1.5). Then the mean field free energy $\mathcal{F}$, defined in (1.14), has a unique minimizer m_∞. The minimizer m_∞ is also the unique solution to the first-order equation (1.15) and an invariant measure to the MFL dynamics (1.12).*

Proof. Recall that $\mathcal{F}(m) = F(m) + H(m)$ where the absolute entropy $H(m)$ is well defined for $m \in \mathcal{P}_2$ and has value in $(-\infty, +\infty]$ thanks to the decomposition

$$\begin{aligned} H(m) &= \int \log m(x) m(x)\,\mathrm{d}x \\ &= \int \log \frac{m(x)}{(2\pi)^{-d/2} e^{-x^2/2}} m(x)\,\mathrm{d}x + \int \left(\log(2\pi)^{-d/2} - \frac{x^2}{2} \right) m(x)\,\mathrm{d}x. \end{aligned} \tag{1.45}$$

The first term, which is the relative entropy between m and a normalized Gaussian, is always nonnegative and the second term is finite. Moreover the free energy $\mathcal{F}$ satisfies

$$\begin{aligned} \mathcal{F}(m) - F(m_0) &\geqslant \int \frac{\delta F}{\delta m}(m_0, x)(m - m_0)(\mathrm{d}x) + H(m) \\ &= -\int \log \hat{m}_0(x)(m - m_0)(\mathrm{d}x) + H(m) = H(m|\hat{m}_0) + \int \log \hat{m}_0(x) m_0(\mathrm{d}x) \end{aligned} \tag{1.46}$$

for all $m, m_0 \in \mathcal{P}_2$ such that m_0 has finite entropy. Since the LSI (1.4) implies the T_2 inequality (1.9), the functional $\mathcal{F}$ has $\mathcal{P}_2$-coercivity:

$$\rho W_2^2(m, \hat{m}_0) \leqslant H(m|\hat{m}_0) \leqslant \mathcal{F}(m) - \int \log \hat{m}_0(x) m_0(\mathrm{d}x) - F(m_0).$$

The conditions (1.2), (1.5) imply also the $\mathcal{P}_2$-lower-continuity of F: if $(m_n)_{n \in \mathbb{N}}$ is a sequence convergent to m in the weak topology of $\mathcal{P}_2$, then we have

$$\begin{aligned} &\liminf_n F(m_n) - F(m) \\ &\quad \geqslant \liminf_n \int \frac{\delta F}{\delta m}(m, x)(m_n - m)(\mathrm{d}x) \\ &\quad = \liminf_n \int \left(\frac{\delta F}{\delta m}(m, x) - \frac{\delta F}{\delta m}(m, 0) \right)(m_n - m)(\mathrm{d}x) \\ &\quad \geqslant \liminf_n \int \left(D_m F(m, 0) \cdot x - \frac{M_{mx}^F}{2}|x|^2 \right)(m_n - m)(\mathrm{d}x) \\ &\quad = 0. \end{aligned}$$

Here the second inequality follows from Taylor's formula and M_{mx}^F denotes the constant in the condition (1.5). The entropy H is also $\mathcal{P}_2$-lower-semicontinuous by the previous decomposition (1.45). The free energy $\mathcal{F}$ is then lower-bounded, coercive, lower-semicontinuous and convex, so there exists unique minimizer in $\mathcal{P}_2$ which we denote by m_∞.

Now we show the equivalence between the minimizing property of the free energy $\mathcal{F}$ and the first-order condition (1.15). If m_0 satisfies (1.15) then $\hat{m}_0 = m_0$ and from (1.46) we deduce $\mathcal{F}(m) \geqslant \mathcal{F}(m_0)$ for all $m \in \mathcal{P}_2$, i.e. m_0 is the minimizer of $\mathcal{F}$. For the reverse implication we refer readers to the necessary part of the proof of [117, Proposition 2.5].

Finally since m_∞ satisfies (1.15) we have

$$\Delta m_\infty + \nabla\cdot\big(D_m F(m_\infty,x)m_\infty\big) = \nabla\cdot\left(m_\infty\nabla\left(\frac{\delta F}{\delta m}(m_\infty,x)+\log m_\infty\right)\right) = 0,$$

and m_∞ is invariant to (1.12). □

Remark 1.35. We will establish the uniqueness of the invariant measure of the MFL in Corollary 1.39 after deriving the free energy dissipation formula (1.48).

1.4.2 Proof of Theorem 1.4

First we recall the definition of AC^2 curves in [4].

Definition 1.36. Let (X,d) be a complete metric space and $x:[a,b]\to X$ be a continuous mapping. We say x is *absolutely continuous* (a.c.) and write $x\in AC\big([a,b];(X,d)\big)$ if there exists $m\in L^1([a,b])$ such that

$$\forall a\leqslant s<t\leqslant b,\qquad d\big(x(s),x(t)\big)\leqslant\int_s^t m(u)du.$$

We say $x\in AC^2\big([a,b];(X,d)\big)$ if additionally $m\in L^2([a,b])$. For a globally defined curve $x:[t_0,+\infty)\to X$ we say x belongs to the class AC^2_{loc} and denote $x\in AC^2_{\mathrm{loc}}\big([t_0,+\infty;(X,d)\big)$, if $x\in AC^2_{\mathrm{loc}}\big([t_0,T];(X,d)\big)$ for every $T\geqslant t_0$.

Now we state the wellposedness and regularity result.

Proposition 1.37 (Existence, uniqueness and regularity of MFL)**.** *Assume F satisfies* (1.3) *and* (1.5)*. Then*

1. *for all $m_0\in\mathcal{P}_2(\mathbb{R}^d)$ there exists a unique continuous flow $m:[0,+\infty)\to\mathcal{P}_2(\mathbb{R}^d)$ solving weakly the Fokker–Planck equation* (1.12)*;*
2. *moreover, this solution has density and finite entropy for positive time:*
$$\forall t>0,\qquad \int|\log m_t(x)|m_t(x)\,\mathrm{d}x<+\infty;$$
3. *if additionally m_{t_0} has finite entropy for some $t_0\geqslant 0$, then the integral*
$$\int_{t_0}^t\int\frac{|\nabla m_s(x)|^2}{m_s(x)}\,\mathrm{d}x\,\mathrm{d}s \tag{1.47}$$
is finite for every $t\geqslant t_0$; therefore $(m_t)_{t\geqslant t_0}\in AC^2_{\mathrm{loc}}\big([t_0,+\infty);(\mathcal{P}_2,W_2)\big)$ and has tangent vector $v_t(x)=-D_mF(m_t,x)-\nabla\log m_t(x)$ for $t\geqslant t_0$ a.e. in the sense of [4, Proposition 8.4.5].

Due to the technical nature of this proposition its proof is postponed to Appendix A.1. Using the results of Proposition 1.37 and applying the formalism of [4], we establish the free energy dissipation formula, which is crucial to our studies on the dynamics of gradient flow.

Proposition 1.38 (Energy dissipation)**.** *Assume F satisfies (1.3) and (1.5). If m_{t_0} is a measure of finite entropy and finite second moment for some $t_0 \geqslant 0$, then the free energy $\mathcal{F}$, defined in (1.14), is absolutely continuous along the flow $(m_t)_{t \geqslant t_0}$ constructed in Proposition 1.37. Moreover it has derivative*

$$\frac{\mathrm{d}\mathcal{F}(m_t)}{\mathrm{d}t} = -\int |D_m F(m_t, x) + \nabla \log m_t(x)|^2 m_t(\mathrm{d}x), \qquad \text{for } t \geqslant t_0 \text{ a.e.} \quad (1.48)$$

Proof. We will apply the chain rule result of [4, Proposition 10.3.18] and we verify its conditions, namely, the differentiability of the free energy $\mathcal{F} = F + H$ and of the flow of measures m_t. Firstly under the conditions (1.3), (1.5), we can apply the argument of [56, Lemma A.2] to show that $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is $-\lambda$-geodesically-convex for some $\lambda > 0$ and it has differential $D_m F(m_t, \cdot)$ at m_t. Secondly the entropy $H : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is also 0-geodesically-convex by the result of [4, Proposition 9.3.9] and for $t \geqslant t_0$ a.e. has subdifferential $\nabla \log m_t$ at m_t by [4, Theorem 10.4.6], thanks to the regularity bounds in the previous Proposition 1.37. Hence the free energy $\mathcal{F} = F + H$ is $-\lambda$-geodesically-convex and has differential $D_m F(m_t, \cdot) + \nabla \log m_t$ at m_t. For the flow of measures m_t we have already obtained its AC^2-regularity in the previous proposition and its tangent vector reads $v_t = -D_m F(m_t, \cdot) - \nabla \log m_t$ at m_t for $t \geqslant t_0$ a.e. Then we can apply the chain rule to obtain the absolute continuity of $t \mapsto \mathcal{F}(m_t)$ and

$$\forall T > t_0, \quad \mathcal{F}(m_T) - \mathcal{F}(m_{t_0}) = \int_{t_0}^{T} \big(D_m F(m_t, x) + \nabla \log m_t(x)\big) \cdot v_t(x) m_t(\mathrm{d}x)\, \mathrm{d}t$$

which is the desired result. □

Corollary 1.39 (Uniqueness of the invariant measure)**.** *Under (1.2), (1.3), (1.4) and (1.5), there exists a unique invariant measure in $\mathcal{P}_2(\mathbb{R}^d)$ to the mean field dynamics (1.12).*

Proof. The existence part is already shown in Proposition 1.34. Let $m_* \in \mathcal{P}_2(\mathbb{R}^d)$ be an invariant measure. We let the initial condition m_0 be equal to m_* and construct according to Proposition 1.37 the MFL solution $(m_t)_{t \geqslant 0}$. By the invariance of m_* we have $m_t = m_*$ for all $t \geqslant 0$, so m_* must have density and finite entropy. We then apply the energy dissipation formula (1.48) and obtain

$$\text{for } x \in \mathbb{R}^d \text{ a.e.}, \qquad D_m F(m_*, x) + \nabla \log m_*(x) = 0.$$

Integrating this equation, we obtain m_* solves the first-order condition (1.15) which has unique solution by Proposition 1.34. □

Now we show the close relation between the free energy and the relative entropies.

Lemma 1.40 (Entropy sandwich)**.** *Assume F satisfies (1.2), (1.3), (1.4) and (1.5). Then for every $m \in \mathcal{P}_2(\mathbb{R}^d)$ we have*

$$\begin{aligned} H(m|m_\infty) \leqslant \mathcal{F}(m) - \mathcal{F}(m_\infty) \leqslant H(m|\hat{m}) \\ \leqslant \left(1 + \frac{M^F_{mm}}{\rho} + \frac{(M^F_{mm})^2}{2\rho^2}\right) H(m|m_\infty). \end{aligned} \quad (1.49)$$

Proof. The first two inequalities are proved in [56, Lemma 3.4]. We show the rightmost one. Recall that $Z(\hat{m})$ is the normalization constant defined in (1.7). We have

$$\begin{aligned}H(m|\hat{m}) - H(m|m_\infty) &= \int \biggl(\log\frac{m}{\hat{m}} - \log\frac{m}{m_\infty}\biggr)m = \int \log\frac{m_\infty}{\hat{m}}m\\&= \int\biggl(\frac{\delta F}{\delta m}(m,x) - \frac{\delta F}{\delta m}(m_\infty,x)\biggr)m(x)\,\mathrm{d}x + \log Z(\hat{m}) - \log Z(m_\infty).\end{aligned}$$

By Jensen's inequality, the difference between $\delta \coloneqq \log Z(\hat{m}) - \log Z(\hat{m}_\infty)$ satisfies

$$\begin{aligned}\delta &= \log Z(\hat{m}) - \log \int \exp\biggl(-\frac{\delta F}{\delta m}(m_\infty,x)\biggr)\mathrm{d}x\\&= \log Z(\hat{m}) - \log \int \exp\biggl(-\frac{\delta F}{\delta m}(m_\infty,x) - \log\hat{m}(x)\biggr)\hat{m}(x)\,\mathrm{d}x\\&\leqslant \log Z(\hat{m}) + \int\biggl(\frac{\delta F}{\delta m}(m_\infty,x) + \log\hat{m}(x)\biggr)\hat{m}(x)\,\mathrm{d}x\\&\leqslant \log Z(\hat{m}) + \int\biggl(\frac{\delta F}{\delta m}(m_\infty,x) - \frac{\delta F}{\delta m}(m,x) - \log Z(\hat{m})\biggr)\hat{m}(x)\,\mathrm{d}x\\&= \int\biggl(\frac{\delta F}{\delta m}(m_\infty,x) - \frac{\delta F}{\delta m}(m,x)\biggr)\hat{m}(x)\,\mathrm{d}x.\end{aligned}$$

Then we have by Kantorovich duality and W_1-Lipschitzianity in (1.3)

$$\begin{aligned}H(m|\hat{m}) - H(m|m_\infty) &\leqslant \int\biggl(\frac{\delta F}{\delta m}(m,x) - \frac{\delta F}{\delta m}(m_\infty,x)\biggr)\bigl(m(x) - \hat{m}(x)\bigr)\,\mathrm{d}x\\&\leqslant \|D_mF(m,x) - D_mF(m_\infty,x)\|_\infty W_1(m,\hat{m})\\&\leqslant M^F_{mm}W_1(m,m_\infty)W_1(m,\hat{m})\\&\leqslant M^F_{mm}W_1(m,m_\infty)\bigl(W_1(m,m_\infty) + W_1(\hat{m},m_\infty)\bigr).\end{aligned}$$

Note that, for the first term in the bracket above, we have

$$W_1(m,m_\infty) \leqslant W_2(m,m_\infty) \leqslant \sqrt{\rho^{-1}H(m|m_\infty)}$$

by the T_2 and log-Sobolev inequalities, (1.9), (1.4), and for the second term, we have

$$\begin{aligned}W_1^2(\hat{m},m_\infty) \leqslant W_2^2(\hat{m},m_\infty) \leqslant \frac{1}{\rho}H(\hat{m}|m_\infty) &\leqslant \frac{1}{4\rho^2}\int\biggl|\nabla\log\frac{\hat{m}}{m_\infty}\biggr|^2\hat{m}\\&= \frac{1}{4\rho^2}\int|D_mF(m,x) - D_mF(m_\infty,x)|^2\hat{m}(x)\,\mathrm{d}x\\&\leqslant \frac{(M^F_{mm})^2}{4\rho^2}W_1^2(m,m_\infty) \leqslant \frac{(M^F_{mm})^2}{4\rho^3}H(m|m_\infty),\end{aligned}$$

which concludes. □

The proof of Theorem 1.4 is nothing but the combination of the previous two results.

Proof of Theorem 1.4. By Proposition 1.38 we have

$$\frac{d\mathcal{F}(m_t)}{dt} = -\int |D_m F(m_t,x) + \nabla \log m_t(x)|^2 m_t(\mathrm{d}x) = -I(m_t|\hat{m}_t)$$
$$\leqslant -4\rho H(m_t|\hat{m}_t) \leqslant -4\rho\big(\mathcal{F}(m_t) - \mathcal{F}(m_\infty)\big), \qquad \text{for } t \geqslant t_0 \text{ a.e.}$$

The first inequality is due to the uniform log-Sobolev inequality (1.4) and the second to the entropy sandwich (1.49). The second inequality in (1.18) is then obtained by Grönwall's lemma, and the first inequality has already been proved in Lemma 1.40. □

1.4.3 L^2-convergence and hypercontractivity

Standard algebra

We first work on dense set of sufficiently regular functions that will be necessary our proofs.

For notational simplicity, define $b_t(x) \coloneqq -D_m F(m_t,x)$, $b_\infty(x) \coloneqq -D_m F(m_\infty,x)$ and recall that $h_t(x) \coloneqq \frac{dm_t}{dm_\infty}(x)$. The relative density h_t then solves

$$\partial_t h = \Delta h + (2b_\infty - b_t)\cdot\nabla h - \big(\nabla\cdot(b_t - b_\infty) + (b_t - b_\infty)\cdot b_\infty\big)h. \qquad (1.50)$$

In this subsection we will fix the flow of measures m_t to be that constructed in Proposition 1.37 and let h change independently from m_t. We will also only consider solutions in $L^\infty\big([t_0,T];L^1(m_\infty)\big)$ with initial value $h_{t_0} \in L^1(m_\infty)$ to the evolution equation (1.50) (in the sense of [22, (6.1.3)]). We then know that the solution is then unique by applying [22, Theorem 9.6.3] to hm_∞.

Definition 1.41 (Standard algebra)**.** The *standard algebra* $\mathcal{A}_+$ is the set of positive and C^2 functions $h : \mathbb{R}^d \to (0,\infty)$ satisfying the following conditions:

- there exists a constant $M > 0$ such that for every $x \in \mathbb{R}^d$, $|\log h(x)| \leqslant M(1+|x|)$;
- for $k = 1$, 2, there exist constants $M_k > 0$ such that for every $x \in \mathbb{R}^d$, $|\nabla^k h(x)| \leqslant \exp\big(M_k(1+|x|)\big)$.

For a collection of functions $(h_i)_{i\in I}$ we say that $h_i \in \mathcal{A}_+$ uniformly for $i \in I$ or $(h_i)_{i\in I} \subset \mathcal{A}_+$ uniformly, if there exist constants M, M_1, M_2 such that the previous bounds holds for all h_i, $i \in I$.

Remark 1.42. The word "standard algebra" is the terminology in [6]. Readers may have noticed $\mathcal{A}_+$ is not an algebra in the usual sense, as it contains only positive functions and is not closed under scalar multiplication by -1. To remedy this we can define $\mathcal{A} = \mathcal{A}_+ - \mathcal{A}_+$ and $\mathcal{A}$ is truely an algebra. We introduce this unusual set of functions in order to do L^p-computations for $p < 1$.

Then we can state the density and stability of $\mathcal{A}_+$.

Proposition 1.43 (Density of $\mathcal{A}_+$)**.** *Let $p \geqslant 1$, $q < 1$, $h : \mathbb{R}^d \to [0,+\infty]$ be a measurable function and μ be a probability measure on $\mathbb{R}^d$ having a density with respect to the Lebesgue measure. If $h \in L^p(\mu)$, then there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $\mathcal{A}_+$ such that $h_n \to h$ in $L^p(\mu)$; if $h \in L^q(\mu)$, then there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $\mathcal{A}_+$ such that $\|h_n\|_q \to \|h\|_q$; and if $h \in L^p \cap L^q(\mu)$, then the sequence in $\mathcal{A}_+$ can be chosen such that both convergences hold.*

Proposition 1.44 (Stability of $\mathcal{A}_+$ under flow)**.** *Assume F satisfies* (1.2), (1.3), (1.4), (1.5) *and* (1.6). *For every $t_0 \geqslant 0$ and $h' \in \mathcal{A}_+$, there exists a solution $h : [t_0, +\infty) \to \mathcal{A}_+$ to* (1.50) *with initial value $h(t_0, \cdot) = h'$. Moreover the temporal weak derivative $\partial_t h$ exists and h_t belongs to $\mathcal{A}_+$ locally uniformly, i.e., $(h_t)_{t\in K} \subset \mathcal{A}_+$ uniformly for every compact subset $K \subset [t_0, +\infty)$.*

The proofs of Propositions 1.43 and 1.44 are postponed to Appendix A.1 due to their technical nature.

Proof of Proposition 1.6

First, by working in $\mathcal{A}_+$, we obtain the following L^p-norm growth formula.

Proposition 1.45 (L^p-norm growth)**.** *Assume F satisfies* (1.2), (1.3), (1.4), (1.5) *and* (1.6). *Let $p \neq 0$ and $h : [a, b] \to \mathcal{A}_+$ be a solution to the evolution* (1.50). *Then the growth of p-norm $t \mapsto \int h_t(x)^p m_\infty(\mathrm{d}x)$ is absolutely continuous and has derivative*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int h_t(x)^p m_\infty(\mathrm{d}x) = p(p-1)\bigg(-\int h_t(x)^{p-2}|\nabla h_t(x)|^2 m_\infty(\mathrm{d}x) + \int h_t(x)^{p-1}\nabla h_t(x)\cdot\big(b_t(x)-b_\infty(x)\big)m_\infty(\mathrm{d}x)\bigg) \tag{1.51}$$

for $t \in [a, b]$ a.e.

Proof. We first suppose $t \mapsto h(t, x)$ is C^1 instead of only absolutely continuous. Notice that the evolution equation (1.50) of h can be rewritten as

$$\partial_t h = (\Delta + b_\infty \cdot \nabla)h - (b_t - b_\infty)\cdot\nabla h - \frac{\nabla\cdot\big(m_\infty(b_t - b_\infty)\big)}{m_\infty}h,$$

where the first term corresponds to the symmetric operator $\Delta + b_\infty \cdot \nabla$ in $L^2(m_\infty)$. We then have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int h_t(x)^p m_\infty(\mathrm{d}x)\\
&\quad= p\int h_t(x)^{p-1}\big(\Delta + b_\infty(x)\cdot\nabla\big)h_t(x)m_\infty(\mathrm{d}x)\\
&\qquad\quad - p\int h_t(x)^{p-1}\big(b_t(x)-b_\infty(x)\big)\cdot\nabla h_t(x)m_\infty(\mathrm{d}x)\\
&\qquad\quad - p\int \nabla\cdot\big(m_\infty(b_t-b_\infty)\big)(x)h_t(x)^p\,\mathrm{d}x\\
&\quad= -p(p-1)\int h_t(x)^{p-2}|\nabla h_t(x)|^2 m_\infty(\mathrm{d}x)\\
&\qquad\quad - p\int h_t(x)^{p-1}\big(b_t(x)-b_\infty(x)\big)\cdot\nabla h_t(x)m_\infty(\mathrm{d}x)\\
&\qquad\quad + p\int \nabla h_t(x)^p\cdot\big(b_t(x)-b_\infty(x)\big)m_\infty(\mathrm{d}x)\\
&\quad= p(p-1)\bigg(-\int h_t(x)^{p-2}|\nabla h_t(x)|^2 m_\infty(\mathrm{d}x)\\
&\qquad\qquad\qquad + \int h_t(x)^{p-1}\nabla h_t(x)\cdot\big(b_t(x)-b_\infty(x)\big)m_\infty(\mathrm{d}x)\bigg).
\end{aligned}$$

We can justify the first equality by the dominated convergence theorem and the two integrations by parts in the second one by an approximating sequence of functions, thanks to the fact that $h_t \in \mathcal{A}_+$ locally uniformly.

Then, for the general case where $t \mapsto h_t(x)$ is only absolutely continuous, thanks to the fact that h_t belongs to $\mathcal{A}_+$ locally uniformly, we have for every s, $t \in [a, b]$ with $s \leqslant t$,

$$\int h_t(x)^p m_\infty(\mathrm{d}x) - \int h_s(x)^p m_\infty(\mathrm{d}x) = p \int_s^t \int h_u(x)^{p-1} \partial_u h_u(x) m_\infty(\mathrm{d}x)\,\mathrm{d}u,$$

where $\partial_u h_u(x)$ is the weak derivative that exists only a.e. Then we plug in the evolution equation (1.50) and compute as before. $\square$

Remark 1.46. By dividing (1.51) by $p-1$ and taking the limit $p \to 1$, one formally obtains

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int h_t(x)\log h_t(x) m_\infty(\mathrm{d}x) = &-\int \frac{|\nabla h_t(x)|^2}{h_t(x)} m_\infty(\mathrm{d}x) \\ &+ \int \nabla h_t(x)\cdot\big(b_t(x) - b_\infty(x)\big) m_\infty(\mathrm{d}x). \quad (1.52)\end{aligned}$$

This entropy growth formula is one of the key ingredients of the method of Jabin and Wang [124] and has also been used in [98]. A weak version of this formula under weak regularity of b has been rigorously proved in the Appendix A of the first arXiv version of [142]. In our case, the formula can be first rigorously proved for h taking value in $\mathcal{A}_+$, as is done in the proposition above, and then we treat the general case by the density of $\mathcal{A}_+$.

The L^p-norm growth formula implies the existence of a strongly continuous semigroup in $L^p(m_\infty)$ for all $p \in [1, +\infty)$.

Corollary 1.47 (L^p-continuity of flow)**.** *Under the hypotheses of Proposition 1.45, for every $p \geqslant 1$ and every $a \leqslant s \leqslant t \leqslant b$ there exists a constant $C_{s,t,p} > 0$ such that*

$$\int h_t(x)^p m_\infty(\mathrm{d}x) \leqslant C_{s,t,p} \int h_s(x)^p m_\infty(\mathrm{d}x)$$

holds for every solutions to (1.50) *in $\mathcal{A}_+$. Therefore the evolution equation* (1.50) *determines a strongly continuous (and positive) semigroup $(P_s^t)_{s\leqslant t}$ in $L_+^p(m_\infty)$ for $p \in [1, +\infty)$.*

Proof. For $h_s \in \mathcal{A}_+$ define $h_t = h(t, \cdot) \in \mathcal{A}_+$ where h is the unique solution of (1.50) in $\mathcal{A}_+$. The mapping $h_s \mapsto h_t$ is linear (when the multiplying scalar is positive). For $p \geqslant 1$, the growth of L^p-norm satisfies

$$\begin{aligned}\frac{d}{du}\int h_u(x)^p m_\infty(\mathrm{d}x) &\leqslant \frac{p(p-1)}{4}\int h_u(x)^p |b_u(x) - b_\infty(x)|^2 m_\infty(\mathrm{d}x) \\ &\leqslant \frac{p(p-1)}{4}(M_{mm}^F)^2 W_1^2(m_u, m_\infty)\int h_u(x)^p m_\infty(\mathrm{d}x)\end{aligned}$$

for $u \in [s, t]$ a.e., by Proposition 1.45 and by Cauchy–Schwarz inequality The existence of the stated constant $C_{s,t,p}$ then follows from an application of Grönwall's lemma. For $p \geqslant 1$, the mapping $h_s \mapsto h_t =: P_s^t h_s$ extends uniquely to a continuous linear one by the density of $\mathcal{A}_+$ in $L^p_+(m_\infty)$. By the dominated convergence theorem we have $\lim_{t\to s}\int |h_t(x)-h_s(x)|^p m_\infty(\mathrm{d}x) = 0$ when $h_s\in\mathcal{A}_+$, using the fact that $(h_u)_{u\in[s,t]}\subset\mathcal{A}_+$ uniformly. This property extends to general $h_s\in L^p_+(m_\infty)$ by the density in Proposition 1.43. Hence P^t_s is a strongly continuous semigroup on $L^p_+(m_\infty)$. To recover the usual definition of strongly continuous semigroup we note that $L^p = L^p_+ - L^p_+$ and define $P^t_s h \coloneqq P^t_s h_+ - P^t_s h_-$ for $h\in L^p(m_\infty)$. ☐

Proof of Proposition 1.6. First suppose $h_{t_0}\in\mathcal{A}_+$. Thanks to Proposition 1.45 with $p=2$, we have

$$\begin{aligned}
&\frac{d}{dt}\int h(x)_t^2 m_\infty(\mathrm{d}x)\\
&\quad= -2\int |\nabla h_t(x)|^2 m_\infty(\mathrm{d}x) + 2\int h_t(x)\nabla h_t(x)\cdot\big(b_t(x)-b_\infty(x)\big)m_\infty(\mathrm{d}x)\\
&\quad\leqslant -2(1-\varepsilon)\int |\nabla h_t(x)|^2 m_\infty(\mathrm{d}x) + \frac{1}{2\varepsilon}\int h_t(x)^2|b_t(x)-b_\infty(x)|^2 m_\infty(\mathrm{d}x)\\
&\quad\leqslant -4(1-\varepsilon)\rho\bigg(\int h_t^2(x)m_\infty(\mathrm{d}x)-1\bigg) + \frac{(M^F_{mm})^2}{2\varepsilon}W_1^2(m_t,m_\infty)\|h_t\|_2^2\\
&\quad= -4(1-\varepsilon)\rho\|h_t-1\|_2^2 + \frac{(M^F_{mm})^2}{2\varepsilon}W_1^2(m_t,m_\infty)\|h_t\|_2^2,
\end{aligned}$$

where we first use the Cauchy–Schwarz inequality before applying the Poincaré inequality (1.8) satisfied by m_∞ and the Lipschitz bound on $|b_t(x)-b_\infty(x)| = |D_mF(m_t,x)-D_mF(m_\infty,x)|$. By the T_2 inequality (1.9) we have $W_1^2(m_t,m_\infty)\leqslant W_2^2(m_t,m_\infty)\leqslant\rho^{-1}H(m_t|m_\infty)$. Thanks to Lemma 1.40 and Theorem 1.4 we have

$$\begin{aligned}
H(m_t|m_\infty) \leqslant \mathcal{F}(m_t) - F(m_\infty) &\leqslant e^{-4\rho(t-t_0)}(\mathcal{F}(m_{t_0})-\mathcal{F}(m_\infty))\\
&\leqslant \bigg(1+\frac{M^F_{mm}}{\rho}+\frac{(M^F_{mm})^2}{2\rho^2}\bigg)e^{-4\rho(t-t_0)}H(m_{t_0}|m_\infty).
\end{aligned}$$

Finally note that the relative entropy satisfies, for $p>1$,

$$H(m_{t_0}|m_\infty)\leqslant \log\|h_{t_0}\|_p^{p/(p-1)} \tag{1.53}$$

since by Jensen's inequality we have

$$\exp\bigg(\int\log\big(h_{t_0}^{p-1}\big)\,\mathrm{d}m_{t_0}\bigg)\leqslant\int h_{t_0}^{p-1}\,\mathrm{d}m_{t_0} = \int h_{t_0}^p\,\mathrm{d}m_\infty.$$

Chaining up the previous three inequalities we obtain

$$\begin{aligned}
\frac{(M^F_{mm})^2}{2\varepsilon}W_1^2(m_t,m_\infty)\leqslant \frac{(M^F_{mm})^2}{2\varepsilon}&W_2^2(m_t,m_\infty)\\
&\leqslant\frac{\rho\alpha^2}{2\varepsilon}\bigg(1+\alpha+\frac{\alpha^2}{2}\bigg)\log\|h_{t_0}\|_2^2 e^{-4\rho(t-t_0)} \eqqcolon \Delta(t),
\end{aligned}$$

where we define $\alpha\coloneqq M^F_{mm}/\rho$. The decrease of L^2-norm then satisfies

$$\frac{d}{dt}\|h_t\|_2^2\leqslant -\big(4\rho'-\Delta(t)\big)\|h_t-1\|_2^2+\Delta(t)$$

with $\rho' \coloneqq (1-\varepsilon)\rho$. Thanks to Grönwall's lemma and the fact that $\int_s^{+\infty} \Delta(u)\,\mathrm{d}u \leqslant \Delta(s)/4\rho$, we obtain

$$\begin{aligned}
&\|h_t - 1\|_2^2 \\
&\quad\leqslant e^{-4\rho'(t-t_0)+\int_{t_0}^t \Delta(s)\,\mathrm{d}s}\|h_{t_0} - 1\|_2^2 + \int_{t_0}^t e^{-4\rho'(t-s)+\int_s^t \Delta(u)du}\Delta(s)\,\mathrm{d}s \\
&\quad\leqslant e^{\Delta(t_0)/4\rho}\biggl(e^{-4\rho'(t-t_0)}\|h_{t_0} - 1\|_2^2 + \int_{t_0}^t e^{-4\rho'(t-s)}\Delta(s)\,\mathrm{d}s\biggr) \\
&\quad\leqslant e^{\Delta(t_0)/4\rho}\biggl(e^{-4\rho'(t-t_0)}\|h_{t_0} - 1\|_2^2 + \Delta(t_0)\int_{t_0}^t e^{-4\rho'(t-s)}e^{-4\rho(s-t_0)}\,\mathrm{d}s\biggr) \\
&\quad\leqslant e^{\Delta(t_0)/4\rho}\biggl(e^{-4\rho'(t-t_0)}\|h_{t_0} - 1\|_2^2 + \frac{\Delta(t_0)}{4(\rho-\rho')}(e^{-4\rho'(t-t_0)} - e^{-4\rho(t-t_0)})\biggr) \\
&\quad\leqslant e^{\Delta(t_0)/4\rho}\biggl(\|h_{t_0} - 1\|_2^2 + \frac{\Delta(t_0)}{4\varepsilon\rho}\biggr)e^{-4\rho'(t-t_0)}.
\end{aligned}$$

For general $h_{t_0} \in L^2(m_\infty)$, we take an approximating sequence $(h_{t_0}^n)_{n\in\mathbb{N}}$ in $\mathcal{A}_+$ such that $h_{t_0}^n \to h_{t_0}$ in $L^2(m_\infty)$ according to Proposition 1.44. We have established that $\|h_t^n - 1\|_2 \leqslant Ce^{-\gamma t}$ where $h_t^n = P_{t_0}^t h_{t_0}^n$. By the continuity shown in Corollary 1.47, we have $h_t^n \to h_t$ in $L^2(m_\infty)$. Therefore, the inequality (1.19) holds for general $h_{t_0} \in L^2(m_\infty)$. □

Proof of Proposition 1.7

Proof of Proposition 1.7. First assume $h_{t_0} \in \mathcal{A}_+$ so that $h_t \in \mathcal{A}_+$ for all $t \geqslant t_0$ and that $h_t \in \mathcal{A}_+$ uniformly on compact sets of $[t_0, +\infty)$ thanks to Proposition 1.44. Define the function $\varphi(t) = \log\|h_t\|_{q(t)}$. In particular, if $q(t) = 0$, then $\varphi(t) = \int \log h_t(x) m_\infty(\mathrm{d}x)$. By the definition of the stable algebra $\mathcal{A}_+$ we know $\varphi(t)$ is well defined for $t \geqslant t_0$. Moreover, it follows from Fubini's theorem that $t \mapsto \varphi(t)$ is absolutely continuous for $t \geqslant t_0$ and its weak derivative reads

$$\begin{aligned}
&\dot\varphi(t) \\
&= \frac{\dot q(t)}{q(t)^2\int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)}\biggl(\int h_t(x)^{q(t)}\log h_t(x)^{q(t)}m_\infty(\mathrm{d}x) \\
&\qquad\qquad - \int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)\log\int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)\biggr) \\
&\quad + \frac{q(t)-1}{\int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)}\biggl(-\int h_t(x)^{q(t)-2}|\nabla h_t(x)|^2 m_\infty(\mathrm{d}x) \\
&\qquad\qquad + \int h_t(x)^{q(t)-1}\nabla h_t(x)\cdot\bigl(b_t(x) - b_\infty(x)\bigr)m_\infty(\mathrm{d}x)\biggr).
\end{aligned}$$

We recognize the term on the first line as the entropy,

$$\begin{aligned}
\int h_t(x)^{q(t)}\log h_t^{q(t)}m_\infty(\mathrm{d}x) - \int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)\log\int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)& \\
= \mathrm{Ent}_{m_\infty}(h_t^{q(t)}),&
\end{aligned}$$

which, by LSI (1.4), has upper bound

$$\mathrm{Ent}_{m_\infty}(h_t^{q(t)}) \leqslant \frac{1}{\rho}\,\mathbb{E}_{m_\infty}\bigl[|\nabla h^{q(t)/2}|^2\bigr] \leqslant \frac{q(t)^2}{4\rho}\int h_t(x)^{q(t)-2}|\nabla h_t(x)|^2 m_\infty(\mathrm{d}x).$$

By Cauchy–Schwarz, the second term on the second line satisfies

$$\int h_t(x)^{q(t)-1}\nabla h_t(x)\cdot\big(b_t(x)-b_\infty(x)\big)m_\infty(\mathrm{d}x)$$
$$\leqslant \varepsilon\int h_t(x)^{q(t)-2}|\nabla h_t(x)|^2 m_\infty(\mathrm{d}x)+\frac{1}{4\varepsilon}\bigg(\int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)\bigg)\|b_t-b_\infty\|_\infty^2$$
$$\leqslant \varepsilon\int h_t(x)^{q(t)-2}|\nabla h_t(x)|^2 m_\infty(\mathrm{d}x)+\frac{(M_{mm}^F)^2W_1^2(m_t,m_\infty)}{4\varepsilon}\int h_t(x)^{q(t)}m_\infty(\mathrm{d}x).$$

Therefore, for $q_0>1$ (so that $q(t)>1,\dot q(t)>0$), we have $\dot\varphi(t)\leqslant\delta(t)$ while for $q_0<1$ (so that $q(t)<1,\dot q(t)<0$) we have $\dot\varphi(t)\geqslant\delta(t)$. To deal with the case $q(t)=0$ we use the continuity of $t\mapsto\varphi(t)$. We have thus shown (1.20) and (1.21) for $h_{t_0}\in\mathcal{A}_+$.

Now consider general $h_{t_0}\in L_+^{q_0}(m_\infty)$. In the case $q_0>1$, we use the density of $\mathcal{A}_+$ (Proposition 1.43) to find a sequence $(h_{t_0}^n)_{n\in\mathbb{N}}$ in $\mathcal{A}_+$ with $h_{t_0}^n\to h_{t_0}$ in L^{q_0}. To each $h_{t_0}^n$ there exists a flow $t\mapsto h_t^n$ in $\mathcal{A}_+$ satisfying (1.20). For $t\geqslant t_0$, we also have $h_t^n\to h_t$ in L^{q_0} by the semigroup property in Corollary 1.47 so that along a subsequence $h_t^n\to h_t$ a.e. By Fatou's lemma we obtain

$$\log\bigg(\int h_t^{q(t)}(x)m_\infty(\mathrm{d}x)\bigg)^{1/q(t)}\leqslant\liminf_{n\to\infty}\bigg(\int h_t^n(x)^{q(t)}m_\infty(\mathrm{d}x)\bigg)^{1/q(t)}$$
$$\leqslant\liminf_{n\to\infty}\log\|h_{t_0}^n\|_{q_0}+\int_{t_0}^t\delta(s)\,\mathrm{d}s=\log\|h_{t_0}\|_{q_0}+\int_{t_0}^t\delta(s)\,\mathrm{d}s.$$

So (1.20) is proved for general $h_{t_0}\in L^{q_0}$. In the case $q_0<1$, we choose again by Proposition 1.43 a sequence $(h_{t_0}^n)_{n\in\mathbb{N}}$ in $\mathcal{A}_+$ such that $h_{t_0}^n\to h_{t_0}$ in L^1 and $\lim_{n\to\infty}\|h_{t_0}^n\|_{q_0}=\|h_{t_0}\|_{q_0}$. By the L^1-continuity, $h_t^n\to h_t$ in L^1 so that along a subsequence $h_t^n\to h_t$ pointwise m_∞-a.e. For $q(t)>0$ we have by Fatou's lemma

$$\liminf_{n\to\infty}\int\big(|h_t^n(x)|+1-|h_t^n(x)|^{q(t)}\big)m_\infty(\mathrm{d}x)\geqslant\int\big(|h_t(x)|+1-|h_t(x)|^{q(t)}\big)m_\infty(\mathrm{d}x).$$

Thus $\limsup_{n\to\infty}\int|h_t^n(x)|^{q(t)}m_\infty(\mathrm{d}x)\leqslant\int|h_t(x)|^{q(t)}m_\infty(\mathrm{d}x)$. So taking $\limsup$ on both sides of the inequality

$$\log\|h_t^n\|_{q(t)}\geqslant\log\|h_{t_0}^n\|_{q_0}+\int_{t_0}^t\delta(s)\,\mathrm{d}s$$

gives us (1.21). For $q(t)<0$ we have directly by Fatou

$$\liminf_{n\to\infty}\int h_t^n(x)^{q(t)}m_\infty(\mathrm{d}x)\geqslant\int h_t(x)^{q(t)}m_\infty(\mathrm{d}x)$$

so that

$$\log\|h_t\|_{q(t)}\geqslant\limsup_{n\to\infty}\log\|h_t^n\|_{q(t)}\geqslant\limsup_{n\to\infty}\log\|h_{t_0}^n\|_{q_0}+\int_{t_0}^t\delta(s)\,\mathrm{d}s$$
$$=\log\|h_{t_0}\|_{q_0}+\int_{t_0}^t\delta(s)\,\mathrm{d}s.$$

To conclude we treat $q(t)=0$ by a continuity argument. Take $\varepsilon'\in(0,\varepsilon)$ and let q' be the solution to $\dot q'=4(1-\varepsilon')\rho(q'-1)$ with $q'(t_0)=q(t_0)=q_0<1$ and

$\delta'(t) = \frac{1}{4\varepsilon'}(q'(t)-1)(M^F_{mm})^2 W_1^2(m_t, m_\infty)$. We have $q'(t) < q(t) = 0$ so that by previous discussions

$$\log\|h_t\|_{q'(t)} \geqslant \log\|h_{t_0}\|_{q_0} + \int_{t_0}^t \delta'(s)\,\mathrm{d}s,$$

whereas $\log\|h_t\|_{q(t)} \geqslant \log\|h_t\|_{q'(t)}$ by the monotonicity of p-norm. We take the limit $\varepsilon' \to \varepsilon$ to obtain (1.21). □

Remark 1.48. The computations are similar to that for the hypercontractivity of a diffusion process whose invariant measure m satisfies a *defective LSI*, i.e. for some $c, \delta \geqslant 0$,

$$\forall f \in C_b^1(\mathbb{R}^d), \qquad \mathrm{Ent}_m(f^2) \leqslant c\,\mathbb{E}_m[|\nabla f|^2] + \delta\,\mathbb{E}_m[|f|^2].$$

See [12, Chapter 5] and [6, Chapter 2] for the link between defective LSI and hypercontractivity.

1.4.4 Proofs of Theorems 1.9 and 1.11

After showing the L^2-convergence and the hypercontractivity, we are finally ready to give the proof of Theorem 1.9.

Proof of Theorem 1.9. We will first use Proposition 1.7 to show that after a finite time h lies in $L^2(m_\infty)$, then use Proposition 1.6 to show that its $L^2(m_\infty)$-norm diminishes exponentially and finally apply Proposition 1.7 again to extend this result to all L^p.

To this end, let $\rho' \in (0, \rho)$ be arbitrary and set $\varepsilon = 1 - \rho'/\rho$. Define $\dot{q}_1(t) = 4(1-\varepsilon)\rho\big(q_1(t) - 1\big)$ with $q_1(0) = p_0$, and we know

$$q_1(s) = (p_0 - 1)\exp\big(4(1-\varepsilon)\rho s\big) + 1.$$

Since $p_0 > 1$, q_1 is exponentially increasing. If $p_0 \in (1, 2)$ we set $t_1 = (4(1-\varepsilon)\rho)^{-1}\log\frac{1}{p_0 - 1}$. This definition ensures that $q_1(t_1) = 2$. Otherwise if $p_0 \geqslant 2$, we simply set $t_1 = 0$. Thus, in both cases, we have

$$t_1 = \frac{1}{4(1-\varepsilon)\rho}\log\frac{1}{(p_0 - 1)\wedge 1}.$$

By the hypercontractivity (1.20) in Proposition 1.7, we have

$$\|h_{t_1}\|_2 \leqslant \exp\bigg(\int_0^{t_1}\delta_1(s)\,\mathrm{d}s\bigg)\|h_0\|_{p_0},$$

where $\delta_1(s) = \frac{1}{4\varepsilon}(q_1(s)-1)(M^F_{mm})^2 W_1^2(m_s, m_\infty)$. On the other hand, we can control the Wasserstein distance $W_1^2(m_s, m_\infty)$ as follows:

$$\begin{aligned}
W_1^2(m_s, m_\infty) \leqslant W_2^2(m_s, m_\infty) &\leqslant \rho^{-1} H(m_s|m_\infty)\\
&\leqslant \rho^{-1}\big(\mathcal{F}(m_s) - \mathcal{F}(m_\infty)\big)\\
&\leqslant \rho^{-1}\big(\mathcal{F}(m_0) - \mathcal{F}(m_\infty)\big)\\
&\leqslant \rho^{-1}\bigg(1 + \frac{M^F_{mm}}{\rho} + \frac{(M^F_{mm})^2}{2\rho^2}\bigg)H(m_0|m_\infty)\\
&\leqslant \rho^{-1}\bigg(1 + \frac{M^F_{mm}}{\rho} + \frac{(M^F_{mm})^2}{2\rho^2}\bigg)\log\|h_0\|_{p_0}^{p_0/(p_0-1)},
\end{aligned}$$

thanks to the T_2 inequality (1.9), Theorem 1.4, Lemma 1.40 and the inequality (1.53). Setting $\alpha \coloneqq M^F_{mm}/\rho$ and $P(\alpha) = \alpha^2 + \alpha^3 + \alpha^4/2$, we get

$$\begin{aligned}
\int_0^{t_1} \delta_1(s)\,\mathrm{d}s &\leqslant \frac{M^F_{mm} p_0}{4\varepsilon(p_0-1)} \biggl(\alpha + \alpha^2 + \frac{\alpha^3}{2}\biggr) \log\|h_0\|_{p_0} \int_0^{t_1} (q_1(s)-1)\,\mathrm{d}s \\
&\leqslant \frac{M^F_{mm} p_0}{4\varepsilon(p_0-1)} \biggl(\alpha + \alpha^2 + \frac{\alpha^3}{2}\biggr) \log\|h_0\|_{p_0} \frac{1}{4(1-\varepsilon)\rho}(2-p_0)_+ \\
&\leqslant \frac{p_0(2-p_0)_+}{16(p_0-1)\varepsilon(1-\varepsilon)} P(\alpha) \log\|h_0\|_{p_0} \eqqcolon M \log\|h_0\|_{p_0}.
\end{aligned}$$

And thus, $\|h_{t_1}\|_2 \leqslant \|h_0\|_{p_0}^{1+M}$. By Proposition 1.6 we know that for all $t \in [t_1, +\infty)$,

$$\begin{aligned}
\|h_t\|_2^2 - 1 &\leqslant \exp\biggl(\frac{P(\alpha)}{4\varepsilon}\log\|h_{t_1}\|_2\biggr)\biggl(\|h_{t_1}\|_2^2 - 1 + \frac{P(\alpha)}{4\varepsilon^2}\log\|h_{t_1}\|_2\biggr) e^{-4(1-\varepsilon)\rho(t-t_1)} \\
&\leqslant \|h_{t_1}\|_2^{P(\alpha)/4\varepsilon}\biggl(1 + \frac{P(\alpha)}{8\varepsilon^2}\biggr)\bigl(\|h_{t_1}\|_2^2 - 1\bigr) e^{-4(1-\varepsilon)\rho(t-t_1)} \\
&\leqslant \biggl(1 + \frac{P(\alpha)}{8\varepsilon^2}\biggr) H_1^{P(\alpha)/4\varepsilon}\bigl(H_1^2 - 1\bigr) e^{-4(1-\varepsilon)\rho(t-t_1)},
\end{aligned}$$

for H_1 being the upper bound of $\|h_{t_1}\|_2$ defined by

$$\log H_1 = \biggl(1 + \frac{p_0(2-p_0)_+ P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\biggr)\log\|h_0\|_{p_0}.$$

Now we define τ_p by

$$\begin{aligned}
\tau_p &= \begin{cases} t_1 + \frac{1}{4(1-\varepsilon)\rho}\log\bigl((p-1)\vee 1\bigr) & \text{if } p > 1, \\ t_1 & \text{if } p \in (0,1) \\ t_1 + \frac{1}{4(1-\varepsilon)\rho}\log\bigl(2(1-p)\bigr) & \text{if } p \leqslant 0 \end{cases} \\
&= \begin{cases} \frac{1}{4(1-\varepsilon)\rho}\log\frac{(p-1)\vee 1}{(p_0-1)\wedge 1} & \text{if } p \geqslant 0, \\ \frac{1}{4(1-\varepsilon)\rho}\log\frac{2(1-p)}{(p_0-1)\wedge 1} & \text{if } p < 0, \end{cases}
\end{aligned}$$

In the case $p > 1$, for $t \geqslant \tau_p$ we set $t_2 = t - (4(1-\varepsilon)\rho)^{-1}\log\bigl((p-1)\vee 1\bigr) \geqslant t_1$ and let q_2 solves $\dot q_2(t) = 4(1-\varepsilon)\rho\bigl(q_2(t)-1\bigr)$ with $q_2(t_2) = 2$. Our choice ensures $q_2(t) = 2 \vee p \geqslant p$. By the hypercontractivity (1.20) we have

$$\|h_t\|_{q_2(t)} \leqslant \exp\biggl(\int_{t_2}^t \delta_2(s)\,\mathrm{d}s\biggr)\|h_{t_2}\|_2,$$

where $\delta_2(s) = \frac{1}{4\varepsilon}\bigl(q_2(s)-1\bigr)(M^F_{mm})^2 W_1^2(m_s, m_\infty)$. The integral of δ_2 can be controlled in the same way as we did to push $p_0 \to 2$ by hypercontractivity:

$$\begin{aligned}
\int_{t_2}^t \delta_2(s)\,\mathrm{d}s &\leqslant \frac{M^F_{mm} p_0}{4\varepsilon(p_0-1)}\biggl(\alpha + \alpha^2 + \frac{\alpha^3}{2}\biggr)\log\|h_0\|_{p_0}\int_{t_2}^t \bigl(q_2(s)-1\bigr)\,\mathrm{d}s \\
&\leqslant \frac{p_0 P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0}(p-2)_+.
\end{aligned}$$

The p-norm then satisfies

$$\begin{aligned}
\log\|h_t\|_p &\leqslant \log\|h_t\|_{q_2(t)} \leqslant \log\|h_{t_2}\|_2 + \frac{p_0(p-2)_+P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0} \\
&\leqslant \frac{1}{2}\big(\|h_{t_2}\|_2^2-1\big) + \frac{p_0(p-2)_+P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0} \\
&\leqslant \frac{1}{2}\bigg(1+\frac{P(\alpha)}{8\varepsilon^2}\bigg)H_1^{P(\alpha)/4\varepsilon}\big(H_1^2-1\big)e^{-4(1-\varepsilon)\rho(t_2-t_1)} \\
&\qquad + \frac{p_0(p-2)_+P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0} \\
&\leqslant \frac{1}{2}\bigg(1+\frac{P(\alpha)}{8\varepsilon^2}\bigg)H_1^{P(\alpha)/4\varepsilon}\big(H_1^2-1\big)e^{-4(1-\varepsilon)\rho(t-\tau_p)} \\
&\qquad + \frac{p_0(p-2)_+P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0}.
\end{aligned}$$

So the upper bound in (1.22) is established. The lower bound follows from the monotonicity of p-norm: we have $\log\|h_t\|_p \geqslant \log\|h_t\|_1 = 0$.

For $p \in (0,1)$, we observe Hölder's inequality

$$\bigg(\int h^p m_\infty\bigg)^{1/(2-p)}\bigg(\int h^2 m_\infty\bigg)^{(1-p)/(2-p)} \geqslant \int h m_\infty = 1,$$

so that for $t \geqslant \tau_p = t_1$ we have $\log\|h_t\|_p \geqslant -\frac{2(1-p)}{p}\log\|h_t\|_2$. Thus we obtain the desired bound by inserting the upper bound for $\|h_t\|_2$.

Finally we treat $p \leqslant 0$. Given $t \geqslant \tau_p$, set $t_3 = t - \big(4(1-\varepsilon)\rho\big)^{-1}\log\big(2(1-p)\big) \geqslant t_1$ and let q_3 solves $\dot q_3(t) = 4(1-\varepsilon)\rho\big(q_3(t)-1\big)$ with $q_3(t_3) = \frac{1}{2}$. Our choice ensures $q_3(t) = p$. Define $\delta_3(s) = \frac{1}{4\varepsilon}\big(q_3(s)-1\big)(M^F_{mm})^2 W_1^2(m_s, m_\infty)$. It satisfies, as done in the previous steps,

$$\int_{t_3}^t \delta_3(s)\,\mathrm{d}s \geqslant -\frac{p_0\big(\frac{1}{2}-p\big)P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0}.$$

We obtain, by the reverse hypercontractivity (1.21),

$$\begin{aligned}
\log\|h_t\|_p &\geqslant \log\|h_{t_3}\|_{\frac{1}{2}} + \int_{t_3}^t \delta_3(s)\,\mathrm{d}s \\
&\geqslant -2\log\|h_{t_3}\|_2 - \frac{p_0\big(\frac{1}{2}-p\big)P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0} \\
&= -\log\big(1+\|h_{t_3}-1\|_2^2\big) - \frac{p_0\big(\frac{1}{2}-p\big)P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0} \\
&\geqslant -\|h_{t_3}-1\|_2^2 - \frac{p_0\big(\frac{1}{2}-p\big)P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0} \\
&\geqslant -\bigg(1+\frac{P(\alpha)}{8\varepsilon^2}\bigg)H_1^{P(\alpha)/4\varepsilon}\big(H_1^2-1\big)e^{-4(1-\varepsilon)\rho(t-t_1)} \\
&\qquad - \frac{p_0\big(\frac{1}{2}-p\big)P(\alpha)}{16(p_0-1)\varepsilon(1-\varepsilon)}\log\|h_0\|_{p_0}.
\end{aligned}$$

Thus, we have established the lower bound in (1.22), for both $p \in (0,1)$ and $p \leqslant 0$. To conclude, we compare again the p-norm with the 1-norm and use the monotonicity. $\square$

To conclude the discussions about the mean field dynamics we show a lemma which uses L^p-norms to control a "cross entropy"-like quantities and use it to obtain the uniform-in-time concentration of measure result in Theorem 1.11. The lemma will also be used in the proof of Theorem 1.14.

Lemma 1.49. *Let μ, $\nu \in \mathcal{P}(\mathbb{R}^d)$ and $h : \mathbb{R}^d \to (0, +\infty)$ be a measurable function. Then for all $p > 0$,*

$$-\frac{1}{p}H(\nu|\mu) + \log\|h\|_{L^{-p}(\mu)} \leqslant \int \log h \,\mathrm{d}\nu \leqslant \frac{1}{p}H(\nu|\mu) + \log\|h\|_{L^p(\mu)}. \tag{1.54}$$

Proof. Let X be a measurable space, μ, ν be probability measures on X and $U : X \to \mathbb{R}$ be a random variable. We have the convex duality inequality (see e.g. [27, Corollary 4.14])

$$\mathbb{E}_\nu[U] \leqslant H(\nu|\mu) + \log \mathbb{E}_\mu[e^U]. \tag{1.55}$$

The right hand side of the inequality is always well defined in $(-\infty, +\infty]$. Take $U = p \log h$. For $p > 0$ we obtain

$$\int \log h \,\mathrm{d}\nu \leqslant \frac{1}{p}H(\nu|\mu) + \frac{1}{p}\log \int e^{p\log h}\,\mathrm{d}\mu = \frac{1}{p}H(\nu|\mu) + \log\|h\|_{L^p(\mu)},$$

and for $p < 0$ we obtain

$$\int \log h \,\mathrm{d}\nu \geqslant \frac{1}{p}H(\nu|\mu) + \log\|h\|_{L^p(\mu)}.$$ $\square$

Proof of Theorem 1.11. Let $f : \mathbb{R}^d \to \mathbb{R}$ be 1-Lipschitz continuous and define for $t \geqslant 0$ the moment-generating function $\psi_{t,f}(\lambda) = \log \mathbb{E}_{m_t} e^{\lambda(f - \mathbb{E}_{m_t} f)}$. The equality in (1.55) can be attained and therefore we have (see also [27, Corollary 4.14])

$$\psi_{t,f}(\lambda) = \sup_{\mu \ll m_t} \lambda(\mathbb{E}_\mu f - \mathbb{E}_{m_t} f) - H(\mu|m_t).$$

For each $\mu \ll m_t$, the first term satisfies

$$\begin{aligned}\mathbb{E}_\mu f - \mathbb{E}_{m_t} f &\leqslant W_1(\mu, m_t) \leqslant W_1(\mu, m_\infty) + W_1(m_t, m_\infty)\\ &\leqslant \sqrt{\frac{1}{\rho}H(\mu|m_\infty)} + W_1(m_t, m_\infty)\end{aligned}$$

by Talagrand's transport inequality (1.9) for m_∞. The second term satisfies

$$\begin{aligned}H(\mu|m_t) &= \int \log\frac{\mathrm{d}\mu}{\mathrm{d}m_t}\,\mathrm{d}\mu = \int \left(\log\frac{\mathrm{d}\mu}{\mathrm{d}m_\infty} - \log h_t\right)\mathrm{d}\mu\\ &= H(\mu|m_\infty) - \int \log h_t \,\mathrm{d}\mu\\ &\geqslant H(\mu|m_\infty) - \frac{1}{p}H(\mu|m_\infty) - \log\|h_t\|_p\end{aligned}$$

for $p > 1$ by the previous Lemma 1.49. Hence for $\lambda \geqslant 0$ the moment-generating function $\psi_{t,f}$ satisfies

$$\begin{aligned}\psi_{t,f}(\lambda) &\leqslant \sup_{\mu \ll m_t} \lambda \sqrt{\frac{1}{\rho} H(\mu|m_\infty)} + \lambda W_1(m_t, m_\infty) - (1 - p^{-1}) H(\mu|m_\infty) + \log\|h_t\|_p \\ &\leqslant \frac{\lambda^2}{4(1 - p^{-1})\rho} + \lambda W_1(m_t, m_\infty) + \log\|h_t\|_p.\end{aligned}$$

For $r, \lambda \geqslant 0$ we have by Markov's inequality

$$\begin{aligned}m_t[f - \mathbb{E} f \geqslant r] &\leqslant e^{-\lambda r} \mathbb{E}_{m_t} e^{\lambda(f - \mathbb{E}_{m_t} f)} \\ &\leqslant \exp\biggl(-\lambda r + \frac{\lambda^2}{4(1 - p^{-1})\rho} + \lambda W_1(m_t, m_\infty) + \log\|h_t\|_p \biggr).\end{aligned}$$

Take $\lambda = 2(1 - p^{-1})\rho$. We obtain

$$\begin{aligned}&m_t[f - \mathbb{E} f \geqslant r] \\ &\leqslant \exp\biggl(-\Bigl(1 - \frac{1}{p}\Bigr)\rho r^2 + 2\Bigl(1 - \frac{1}{p}\Bigr)\rho W_1(m_t, m_\infty) r + \log\|h_t\|_p \biggr).\end{aligned}$$

The bound on $m_t[f - \mathbb{E} f \leqslant -r]$ is obtained by applying the previous inequality to $-f$. Given $\rho' \in (0, \rho)$, find $p > 1$ such that $(1 - p^{-1})\rho = \rho'$. The desired result follows from Theorems 1.4 and 1.9. □

Remark 1.50. Our proof is based on the standard transport method for concentration inequalities and we refer readers to [148, Chapter 6] and [27, Chapter 8] for an introduction to it. In fact, our method allows us to prove a more general perturbative result: if m satisfies a T_1 inequality, $h \in L^p_+(m)$ for $p > 1$ and $\int hm = 1$, then hm also has Gaussian concentration (albeit with a weaker constant).

1.5 Particle system

1.5.1 Proof of Theorem 1.12

Before giving the proof of Theorem 1.12 we first show two lemmas on entropies.

Lemma 1.51 (Information inequalities)**.** *Let $X_1, \ldots, X_N$ be measurable spaces, μ be a probability measure on the product space $X = X_1 \times \cdots \times X_N$ and $\nu = \nu^1 \otimes \cdots \otimes \nu^N$ be a σ-finite measure. Then*

$$\sum_{i=1}^N H(\mu^i|\nu^i) \leqslant H(\mu|\nu) \leqslant \sum_{i=1}^N \int H\Bigl(\mu^{i|-i}(\cdot|\boldsymbol{x}^{-i}) \Big| \nu^i\Bigr) \mu^{-i}(\mathrm{d}\boldsymbol{x}^{-i}). \tag{1.56}$$

Here we set the rightmost term to $+\infty$ if the conditional distribution $\mu^{i|-i}$ does not exist μ^{-i}-a.e.

Proof. The inequality is non-trivial only if $\mu \ll \nu$ and in this case we denote the relative density by $f = \mathrm{d}\mu/\mathrm{d}\nu$. For $I \subset \{1, \ldots, N\}$, we define the conditional densities by

$$f^{I|-I}(\boldsymbol{x}^I|\boldsymbol{x}^{-I}) = \begin{cases} \dfrac{f(\boldsymbol{x}^I, \boldsymbol{x}^{-I})}{\int f(\boldsymbol{x}^I, \boldsymbol{x}^{-I}) \nu^{-I}(\mathrm{d}\boldsymbol{x}^{-I})} & \text{if } \displaystyle\int f(\boldsymbol{x}^I, \boldsymbol{x}^{-I}) \nu^{-I}(\mathrm{d}\boldsymbol{x}^{-I}) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The conditional measures are defined via densities

$$\mu^{I|-I}(\mathrm{d}\boldsymbol{x}^I) = f^{I|-I}(\boldsymbol{x}^I|\boldsymbol{x}^{-I})\nu^I(\mathrm{d}\boldsymbol{x}^I).$$

In particular, we do not need the regularity of the underlying spaces $X_1, \dots, X_N$ in order to apply disintegration theorems. Define $I_i = \{1, \dots, i\}$ for $i = 1, \dots, N$. The relative entropy admits the decomposition

$$H(\mu|\nu) = \sum_{i=1}^{N} \int H\Big(\mu^{i|I_{i-1}}(\cdot|\boldsymbol{x}^{I_{i-1}})\Big|\nu^i\Big)\mu^{I_{i-1}}(\mathrm{d}\boldsymbol{x}^{I_{i-1}}).$$

We conclude by applying Jensen's inequality to the convex mappings $\lambda^i \mapsto H(\lambda^i|\nu^i)$. □

Lemma 1.52. *Assume that F satisfies (1.2) and there exists a measure $m_\infty \in \mathcal{P}_2(\mathbb{R}^d)$ verifying (1.15). Then for all $m^N \in \mathcal{P}_2(\mathbb{R}^{dN})$ of finite entropy, we have*

$$H(m^N|m_\infty^{\otimes N}) \leqslant \mathcal{F}^N(m^N) - N\mathcal{F}(m_\infty). \tag{1.57}$$

Proof. Let $\boldsymbol{X}$ be a random variable distributed as m^N. By the convexity of F we have

$$\begin{aligned}
&\mathcal{F}^N(m^N) - N\mathcal{F}(m_\infty)\\
&\quad= \mathbb{E}[NF(\mu_{\boldsymbol{X}}) - NF(m_\infty)] + H(m^N) - NH(m_\infty)\\
&\quad\geqslant \mathbb{E}\bigg[N\int\frac{\delta F}{\delta m}(m_\infty, x)(\mu_{\boldsymbol{X}} - m_\infty)(\mathrm{d}x)\bigg] + H(m^N) - NH(m_\infty)\\
&\quad= -\mathbb{E}\bigg[N\int \log m_\infty(x)(\mu_{\boldsymbol{X}} - m_\infty)(\mathrm{d}x)\bigg] + H(m^N) - NH(m_\infty)\\
&\quad= -\mathbb{E}\bigg[N\int \log m_\infty(x)\mu_{\boldsymbol{X}}(\mathrm{d}x)\bigg] + H(m^N)\\
&\quad= -\int\sum_{i=1}^{N}\log m_\infty(x^i)m^N(\mathrm{d}\boldsymbol{x}) + H(m^N) = H(m^N|m_\infty^{\otimes N}).
\end{aligned}$$ □

Proof of Theorem 1.12. Let $t_0 \geqslant 0$ be such that m_{t_0} has finite entropy and finite second moment. Since $\nabla_i NF(\mu_{\boldsymbol{x}}) = D_m F(\mu_{\boldsymbol{x}}, x^i)$ corresponds to the drift of (1.11), we recognize the particle system flow of measure m_t^N as a linear Langevin flow with the invariant measure m_∞^N, defined in (1.17). In particular, Proposition 1.38 applied to this dynamics yields

$$\frac{d\mathcal{F}^N(m_t^N)}{dt} = -I(m_t^N|m_\infty^N) \tag{1.58}$$

for $t \geqslant t_0$ a.e. In the following we establish a lower bound of the relative Fisher information $I_t \coloneqq I(m_t^N|m_\infty^N)$ in order to obtain the desired result. We divide the proof into several steps.

Step 1: Regularity of conditional distribution. By the elliptic positivity (see e.g. [22, Theorem 8.2.1]), we know that for all $t > t_0$ and $\boldsymbol{x} \in \mathbb{R}^{dN}$, $m_t^N(\boldsymbol{x}) > 0$ with explicit lower bound. Let $i \in \{1, \dots, N\}$. Define marginal density $m_t^{N,-i}(\boldsymbol{x}^{-i}) = \int m_t^N(\boldsymbol{x})\,\mathrm{d}x^i$. It is strictly positive everywhere by the positivity of m_t^N and is lower semicontinuous (in $\boldsymbol{x}^{-i}$) thanks to the continuity of $\boldsymbol{x} \mapsto m_t^N(\boldsymbol{x})$ and Fatou's lemma. Since Fubini gives $\int m_t^{N,-i}(\boldsymbol{x}^{-i})\,\mathrm{d}\boldsymbol{x}^{-i} = 1$, we have $m_t^{N,-i}(\boldsymbol{x}^{-i}) < +\infty$ everywhere. We are therefore able to define the conditional probability density

$$m_t^{N,i|-i}(x^i|\boldsymbol{x}^{-i}) = \frac{m_t^N(\boldsymbol{x})}{m_t^{N,-i}(\boldsymbol{x}^{-i})} = \frac{m_t^N(\boldsymbol{x})}{\int m_t^N(\boldsymbol{x})\,\mathrm{d}x^i}$$

which has generalized derivative in x^i and is strictly positive everywhere.

Step 2: Decomposing Fisher componentwise. Using the conditional distributions, we can decompose the relative Fisher information by

$$\begin{aligned} I_t &= \int \left| \nabla \log \frac{m_t^N(\boldsymbol{x})}{m_\infty^N(\boldsymbol{x})} \right|^2 m_t^N(\mathrm{d}\boldsymbol{x}) = \mathbb{E}\left[\left| \nabla \log \frac{m_t^N(\boldsymbol{X}_t)}{m_\infty^N(\boldsymbol{X}_t)} \right|^2 \right] \\ &= \sum_{i=1}^N \mathbb{E}\left[\left| \nabla_{x^i} \log \frac{m_t^{N,i|-i}(X_t^i|\boldsymbol{X}_t^{-i}) m_t^{N,-i}(\boldsymbol{X}_t^{-i})}{m_\infty^N(\boldsymbol{X}_t)} \right|^2 \right] \\ &= \sum_{i=1}^N \mathbb{E}\left[\left| \nabla_{x^i} \log \frac{m_t^{N,i|-i}(X_t^i|\boldsymbol{X}_t^{-i})}{m_\infty^N(\boldsymbol{X}_t)} \right|^2 \right] \\ &= \sum_{i=1}^N \mathbb{E}\left[\left| \nabla_{x^i} \log m_t^{N,i|-i}(X_t^i|\boldsymbol{X}_t^{-i}) + D_m F(\mu_{\boldsymbol{X}_t}, X_t^i) \right|^2 \right]. \end{aligned}$$

Step 3: Change of empirical measure and componentwise LSI. We replace the empirical measure $\mu_{\boldsymbol{x}}$ in $D_m F$ by $\mu_{\boldsymbol{x}^{-i}}$. Define $\delta_1^i(\boldsymbol{x}; y) = D_m F(\mu_{\boldsymbol{x}}, y) - D_m F(\mu_{\boldsymbol{x}^{-i}}, y)$. Take $\varepsilon \in (0,1)$. The Fisher information satisfies

$$\begin{aligned} I_t &= \sum_{i=1}^N \mathbb{E}\left[\left| \nabla_{x^i} \log m_t^{N,i|-i}(X_t^i|\boldsymbol{X}_t^{-i}) + D_m F(\mu_{\boldsymbol{X}_t^{-i}}, X_t^i) + \delta_1^i(\boldsymbol{X}_t; X_t^i) \right|^2 \right] \\ &\geqslant \sum_{i=1}^N \mathbb{E}\left[\begin{aligned} (1-\varepsilon) \left| \nabla_{x^i} \log m_t^{N,i|-i}(X_t^i|\boldsymbol{X}_t^{-i}) + D_m F(\mu_{\boldsymbol{X}_t^{-i}}, X_t^i) \right|^2 & \\ - (\varepsilon^{-1} - 1) |\delta_1^i(\boldsymbol{X}_t; X_t^i)|^2 & \end{aligned} \right] \\ &= (1-\varepsilon) \sum_{i=1}^N \mathbb{E}\left[I\left(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i}) \middle| \hat{\mu}_{\boldsymbol{X}_t^{-i}} \right) \right] - (\varepsilon^{-1} - 1) \sum_{i=1}^N \mathbb{E}[|\delta_1^i(\boldsymbol{X}_t; X_t^i)|^2], \end{aligned}$$

where we used the elementary inequality $(a+b)^2 \geqslant (1-\varepsilon)|a|^2 - (\varepsilon^{-1}-1)|b|^2$ and $\hat{\mu}_{\boldsymbol{x}^{-i}}$ is the probability of density proportional to $\exp\bigl(-\frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}^{-i}}, x)\bigr)\,\mathrm{d}x$. Define the first error

$$\Delta_1 \coloneqq \sum_{i=1}^N \mathbb{E}[|\delta_1^i(\boldsymbol{X}_t; X_t^i)|^2] \coloneqq \sum_{i=1}^N \mathbb{E}\bigl[|D_m F(\mu_{\boldsymbol{X}_t}, X_t^i) - D_m F(\mu_{\boldsymbol{X}_t^{-i}}, X_t^i)|^2 \bigr]. \quad (1.59)$$

The previous inequality writes

$$I_t \geqslant (1-\varepsilon) \sum_{i=1}^N \mathbb{E}\left[I\left(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i}) \middle| \hat{\mu}_{\boldsymbol{X}_t^{-i}} \right) \right] - (\varepsilon^{-1} - 1)\Delta_1. \quad (1.60)$$

We apply the uniform log-Sobolev inequality for $\hat{\mu}_{\boldsymbol{X}_t^i}$ and obtain

$$\begin{aligned} &\frac{1}{4\rho} I\left(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i}) \middle| \hat{\mu}_{\boldsymbol{X}_t^{-i}} \right) \geqslant H\left(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i}) \middle| \hat{\mu}_{\boldsymbol{X}_t^{-i}} \right) \\ &= \int \left(\log m_t^{N,i|-i}(x^i|\boldsymbol{X}_t^{-i}) + \frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t^{-i}}, x^i) \right) m_t^{N,i|-i}(\mathrm{d}x^i|\boldsymbol{X}_t^{-i}) + \log Z(\hat{\mu}_{\boldsymbol{X}_t^{-i}}). \end{aligned}$$

Then we apply Jensen's inequality to $\log Z(\hat\mu_{\boldsymbol{x}^{-i}})$ to obtain

$$\log Z(\hat\mu_{\boldsymbol{X}_t^{-i}}) \geqslant -\int \frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t^{-i}}, x^i) m_\infty(\mathrm{d}x^i) - \int m_\infty(x^i) \log m_\infty(x^i)\,\mathrm{d}x^i.$$

Chaining the previous two inequalities and summing over i, we have

$$\begin{aligned}\frac{1}{4\rho}\sum_{i=1}^N I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i})\Big|\hat\mu_{\boldsymbol{X}_t^{-i}}\Big) \geqslant \sum_{i=1}^N \bigg[\int \frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t^{-i}}, x^i) \\ \Big(m_t^{N,i|-i}(\mathrm{d}x^i|\boldsymbol{X}_t^{-i}) - m_\infty(\mathrm{d}x^i)\Big) + H\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i})\Big) - H(m_\infty)\bigg]. \quad (1.61)\end{aligned}$$

Step 4: Another change of empirical measure. We wish to change back $\mu_{\boldsymbol{x}^{-i}} \to \mu_{\boldsymbol{x}}$ in (1.61). Define $\delta_2^i(\boldsymbol{x}; y) \coloneqq \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}^{-i}}, y) - \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}}, y)$ and the second error

$$\Delta_2 \coloneqq \sum_{i=1}^N \int \delta_2^i(\boldsymbol{x}; x^i) m_t^N(\mathrm{d}\boldsymbol{x}) - \sum_{i=1}^N \iint \delta_2^i(\boldsymbol{x}; x') m_\infty(\mathrm{d}x') m_t^N(\mathrm{d}\boldsymbol{x}). \quad (1.62)$$

Then we obtain by taking expectations on both sides of (1.61)

$$\begin{aligned}\frac{1}{4\rho}\sum_{i=1}^N \mathbb{E}\Big[I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i})\Big|\hat\mu_{\boldsymbol{X}_t^{-i}}\Big)\Big] \geqslant N\,\mathbb{E}\bigg[\int \frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t}, y)(\mu_{\boldsymbol{X}_t} - m_\infty)(\mathrm{d}y)\bigg] \\ + \sum_{i=1}^N \mathbb{E}\,H\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i})\Big) - NH(m_\infty) + \Delta_2. \quad (1.63)\end{aligned}$$

Thanks to the convexity of F, the first term satisfies the tangent inequality

$$\begin{aligned}N\,\mathbb{E}\bigg[\int \frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t}, y)(\mu_{\boldsymbol{X}_t} - m_\infty)(\mathrm{d}y)\bigg] \geqslant N\,\mathbb{E}\big[F(\mu_{\boldsymbol{X}_t}) - F(m_\infty)\big] \\ = F^N(m_t^N) - NF(m_\infty). \quad (1.64)\end{aligned}$$

For the second term we apply the information inequality (1.56) to obtain

$$\sum_{i=1}^N \mathbb{E}^{-i}\Big[H\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i})\Big)\Big] \geqslant H(m_t^N).$$

Hence,

$$\begin{aligned}\sum_{i=1}^N \mathbb{E}\Big[I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{X}_t^{-i})\Big|\hat\mu_{\boldsymbol{X}_t^{-i}}\Big)\Big] \\ \geqslant 4\rho\big(F^N(m_t^N) - NF(m_\infty) + H(m_t^N) - NH(m_\infty) + \Delta_2\big).\end{aligned}$$

Using (1.60) and recalling the definition of free energies $\mathcal{F}(m) = F(m) + H(m)$, $\mathcal{F}^N(m^N) = F^N(m^N) + H(m^N)$, we obtain

$$I_t = I(m_t^N|m_\infty^N) \geqslant 4\rho(1-\varepsilon)\big(\mathcal{F}^N(m_t^N) - N\mathcal{F}(m_\infty) + \Delta_2\big) - (\varepsilon^{-1}-1)\Delta_1. \quad (1.65)$$

Step 5: Estimate of the errors Δ_1, Δ_2.. The transport plan between $\mu_{\boldsymbol{x}}$ and $\mu_{\boldsymbol{x}^{-i}}$

$$\pi^i = \frac{1}{N}\sum_{j\neq i}\delta_{(x^j,x^j)} + \frac{1}{N(N-1)}\sum_{j\neq i}\delta_{(x^j,x^i)} \tag{1.66}$$

gives the bound

$$W_1(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}}) \leqslant \frac{1}{N(N-1)}\sum_{j\neq i}|x^j - x^i|.$$

We use this transport plan to bound the errors Δ_1, Δ_2.

Let us treat the first error Δ_1. Since $m \mapsto D_mF(m,x)$ is M^F_{mm}-Lipschitz continuous in W_1 metric, we have

$$|\delta^i_1(\boldsymbol{x};y)| \leqslant M^F_{mm}W_1(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}}) \leqslant \frac{M^F_{mm}}{N(N-1)}\sum_{j=1,j\neq i}^N |x^j - x^i|.$$

Under the L^2-optimal transport plan $\operatorname{Law}\big((X^i_t)^N_{i=1},(\tilde{X}^i_\infty)^N_{i=1}\big) \in \Pi(m^N_t, m^{\otimes N}_\infty)$ we have

$$\begin{aligned}
\Delta_1 &= \sum_{i=1}^N \mathbb{E}[|\delta^i_1(\boldsymbol{X}_t;X^i_t)|^2] \leqslant (M^F_{mm})^2\sum_{i=1}^N \mathbb{E}[W_1^2(\mu_{\boldsymbol{X}_t},\mu_{\boldsymbol{X}_t^{-i}})] \\
&\leqslant \frac{(M^F_{mm})^2}{N(N-1)}\,\mathbb{E}\Bigg[\sum_{\substack{1\leqslant i,j\leqslant N\\ i\neq j}} |X^j_t - X^i_t|^2\Bigg] \\
&\leqslant \frac{3(M^F_{mm})^2}{N(N-1)}\,\mathbb{E}\Bigg[\sum_{\substack{1\leqslant i,j\leqslant N\\ i\neq j}} \Big(|X^i_t - \tilde{X}^i_\infty|^2 + |\tilde{X}^i_\infty - \tilde{X}^j_\infty|^2 + |X^j_t - \tilde{X}^j_\infty|^2\Big)\Bigg] \\
&\leqslant \frac{3(M^F_{mm})^2}{N(N-1)}\Bigg(2(N-1)\,\mathbb{E}\Bigg[\sum_{i=1}^N |X^i_t - \tilde{X}^i_\infty|^2\Bigg] + N(N-1)\,\mathbb{E}[|\tilde{X}^1_\infty - \tilde{X}^2_\infty|^2]\Bigg).
\end{aligned}$$

The first term $\mathbb{E}[\sum_{i=1}^N |X^i_t - \tilde{X}^i_\infty|^2]$ is the Wasserstein distance $W_2^2(m^N_t, m^{\otimes N}_\infty)$, while the second $\mathbb{E}[|\tilde{X}^1_\infty - \tilde{X}^2_\infty|^2]$ equals $2\operatorname{Var} m_\infty$. Hence the first error satisfies the bound

$$\Delta_1 \leqslant 6(M^F_{mm})^2\bigg(\frac{1}{N}W_2^2(m^N_t, m^{\otimes N}_\infty) + \operatorname{Var} m_\infty\bigg). \tag{1.67}$$

Now treat the second error Δ_2. The Lipschitz constant of the mapping $y \mapsto \delta^i_2(\boldsymbol{x};y) = \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}^{-i}},y) - \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}},y)$ is controlled by

$$|\nabla_y\delta^i_2(\boldsymbol{x};y)| = |D_mF(\mu_{\boldsymbol{x}},y) - D_mF(\mu_{\boldsymbol{x}^{-i}},y)| \leqslant M^F_{mm}W_1(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}}).$$

Hence we have

$$|\delta^i_2(\boldsymbol{x};y) - \delta^i_2(\boldsymbol{x};y')| \leqslant M^F_{mm}W_1(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}})|y-y'|.$$

Use Fubini's theorem to first integrate x' in the definition of the second error (1.62) and let $\tilde{X}'_\infty$ be independent from $\boldsymbol{X}_t$. Then we obtain

$$\begin{aligned}
|\Delta_2| &\leqslant \sum_{i=1}^N \int \biggl(\int |\delta_2^i(\boldsymbol{x};x^i) - \delta_2^i(\boldsymbol{x};x')| m_\infty(\mathrm{d}x') \biggr) m_t^N(\mathrm{d}\boldsymbol{x}) \\
&\leqslant \sum_{i=1}^N \iint \frac{M_{mm}^F}{N(N-1)} \sum_{j=1\, j\neq i}^N |x^j - x^i||x' - x^i| m_\infty(\mathrm{d}x') m_t^N(\mathrm{d}\boldsymbol{x}) \\
&= \frac{M_{mm}^F}{N(N-1)} \sum_{\substack{i,j=1\\ i\neq j}}^N \mathbb{E}[|X_t^j - X_t^i||X_t^i - \tilde{X}'_\infty|] \\
&\leqslant \frac{M_{mm}^F}{2N(N-1)} \Biggl(\sum_{\substack{i,j=1\\ i\neq j}}^N \mathbb{E}\,|X_t^i - X_t^j|^2 + (N-1) \sum_{i=1}^N \mathbb{E}\,|X_t^i - \tilde{X}'_\infty|^2 \Biggr).
\end{aligned}$$

Using the same method we used for Δ_1, we control the first term by

$$\sum_{\substack{i,j=1\\ i\neq j}}^N \mathbb{E}\,|X_t^i - X_t^j|^2 \leqslant 6N(N-1) \biggl(\frac{1}{N} W_2^2(m_t^N, m_\infty^{\otimes N}) + \operatorname{Var} m_\infty \biggr).$$

For the second term we work again under the L^2-optimal plan

$$\mathrm{Law}\bigl((X_t^i)_{i=1}^N, (\tilde{X}_\infty^i)_{i=1}^N\bigr) \in \Pi(m_t^N, m_\infty^{\otimes N})$$

and let $\tilde{X}'_\infty$ remain independent from the other variables. We have

$$\begin{aligned}
\sum_{i=1}^N \mathbb{E}\,|X_t^i - \tilde{X}'_\infty|^2 &\leqslant 2 \sum_{i=1}^N \Bigl(\mathbb{E}\,|X_t^i - \tilde{X}_\infty^i|^2 + |\tilde{X}_\infty^i - \tilde{X}'_\infty|^2 \Bigr) \\
&= 2N \biggl(\frac{1}{N} W_2^2(m_t^N, m_\infty^{\otimes N}) + 2 \operatorname{Var} m_\infty \biggr).
\end{aligned}$$

As a result,

$$|\Delta_2| \leqslant M_{mm}^F \biggl(\frac{4}{N} W_2^2(m_t^N, m_\infty^{\otimes N}) + 5 \operatorname{Var} m_\infty \biggr). \tag{1.68}$$

Step 5: Conclusion. Inserting the bounds on the errors (1.67), (1.68) to the lower bound of Fisher information (1.65), we obtain

$$\begin{aligned}
I(m_t^N|m_\infty^N) \geqslant{}& 4\rho(1-\varepsilon)\bigl(\mathcal{F}^N(m_t^N) - N\mathcal{F}(m_\infty)\bigr) \\
&- \bigl(16\rho M_{mm}^F + 6(\varepsilon^{-1}-1)(M_{mm}^F)^2\bigr) \frac{1}{N} W_2^2(m_t^N, m_\infty^{\otimes N}) \\
&- \bigl(20\rho M_{mm}^F + 6(\varepsilon^{-1}-1)(M_{mm}^F)^2\bigr) \operatorname{Var} m_\infty.
\end{aligned}$$

Thanks to the Poincaré inequality (1.8) for $m_\infty = \hat{m}_\infty$, its variance satisfies

$$2\rho \operatorname{Var}_{m_\infty}(x^i) \leqslant \mathbb{E}_{m_\infty}\bigl[|\nabla x^i|^2\bigr] = 1.$$

So $\operatorname{Var} m_\infty = \sum_{i=1}^d \operatorname{Var}_{m_\infty}(x^i) \leqslant d/2\rho$. Using the T_2-transport inequality (1.9) for $m_\infty^{\otimes N}$ and the entropy sandwich Lemma 1.52 we control the transport cost by

$$W_2^2(m_t^N, m_\infty^{\otimes N}) \leqslant \frac{1}{\rho} H(m_t^N|m_\infty^{\otimes N}) \leqslant \frac{1}{\rho}\bigl(\mathcal{F}^N(m_t^N) - N\mathcal{F}(m_t)\bigr).$$

In the end we obtain

$$\begin{aligned}\frac{d\mathcal{F}^N(m_t^N)}{dt} &= -I(m_t^N|m_\infty^N)\\ &\leqslant -\left(4(1-\varepsilon)\rho - \frac{M_{mm}^F}{N}\left(16+6(\varepsilon^{-1}-1)\frac{M_{mm}^F}{\rho}\right)\right)\big(\mathcal{F}^N(m_t^N)-N\mathcal{F}(m_\infty)\big)\\ &\qquad + dM_{mm}^F\left(10+3(\varepsilon^{-1}-1)\frac{M_{mm}^F}{\rho}\right).\end{aligned}$$

We conclude by applying Grönwall's lemma to the differential inequality above and using the entropy inequality of Lemma 1.52. □

Remark 1.53. If the initial condition m_0^N of the particle system is a tensor product $(m_0)^{\otimes N}$, one may expect the (non-uniform) convergence of the free energy $\frac{1}{N}\mathcal{F}(m_t^N) \to \mathcal{F}(m_t)$ for all $t \geqslant 0$. If this is true, one can take the limit $N \to \infty$ to recover the result of Theorem 1.4. However, while the convergence of the regular part $\frac{1}{N}F(m_t^N) \to F(m_t)$ can be expected from the finite-time Wasserstein convergence $\frac{1}{N}\sup_{t\in[0,T]} W_2(m_t^N, m_t^{\otimes N}) \to 0$, the convergence of entropy $H(m_t^N) \to H(m_t^{\otimes N})$ is more difficult to obtain.

Remark 1.54. We used the convexity of F to achieve two things in the proof: (i) the existence of mean field invariant measure m_∞; and (ii) to derive (1.64) and (1.57). Under mild assumptions (i) can also be obtained by a Schauder-type fixed point theorem for the mapping $m \mapsto \hat{m}$, or by finding stationary points of the mean field free energy $\mathcal{F}$. For (ii), if F is only $-\kappa$-semi-convex around m_∞, in the sense that

$$F(m) - F(m_\infty) \geqslant \int \frac{\delta F}{\delta m}(m_\infty, x)(m - m_\infty)(\mathrm{d}x) - \frac{\kappa^2}{2}W_2^2(m, m_\infty),$$

we can expect our method to apply as long as κ is sufficiently small.

1.5.2 Proofs of Theorem 1.14 and Corollary 1.15

Proof of Theorem 1.14. We separate the proof in two parts, each dealing with the finite-time and long-time propagation of chaos respectively. In each part, we shall first control the Wasserstein distance $W_2(m_t^N, m_t^{\otimes N})$ between the particle system and the tensorized mean field system, and then control their relative entropy $H(m_t^N|m_t^{\otimes N})$.

Part 1: Finite-time behavior. We shall use the synchronous coupling method to control the Wasserstein distance between m_t^N and $m_t^{\otimes N}$ and use Girsanov's theorem to control their relative entropy on finite time intervals. This may be considered folklore by specialists and the method of proof has appeared in the end of Chapter 6 of [35]. We, however, include a proof for the sake of self-containedness.

First let us show the bound on the Wasserstein distance $W_2(m_t^N, m_\infty^{\otimes N})$. Recall that $\boldsymbol{X}_t = (X_t^i)_{i=1}^N$ is the solution of the SDE (1.11) with Brownian motions $(W^i)_{i=1}^N$. Let $\tilde{\boldsymbol{X}}_t^i = (\tilde{X}_t^i)_{i=1}^N$ solve

$$\mathrm{d}\tilde{X}_t^i = -D_m F(m_t, \tilde{X}_t^i)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i, \quad i = 1, \dots, N$$

with the initial condition $\mathrm{Law}(\tilde{X}_0^1, \dots, \tilde{X}_t^N) = m_0^{\otimes N}$ and

$$W_2^2(m_0^N, m_t^{\otimes N}) = \sum_{i=1}^N \mathbb{E}\big[|X_0^i - \tilde{X}_t^i|^2\big],$$

i.e., the couple $(\boldsymbol{X}_0^i, \tilde{\boldsymbol{X}}_0^i)$ is distributed as the L^2-optimal transport plan between m_0^N and $m_0^{\otimes N}$. Then, by subtracting the dynamical equations of $\boldsymbol{X}_t$ and $\tilde{\boldsymbol{X}}_t$, we have

$$\begin{aligned}\mathrm{d}\bigg(\sum_{i=1}^N |X_t^i - \tilde{X}_t^i|^2\bigg) &= -2\sum_{i=1}^N (X_t^i - \tilde{X}_t^i)\cdot\big(D_mF(\mu_{\boldsymbol{X}_t}, X_t^i) - D_mF(m_t, \tilde{X}_t^i)\big) \\ &\leqslant \sum_{i=1}^N |X_t^i - \tilde{X}_t^i|^2 + \sum_{i=1}^N |D_mF(\mu_{\boldsymbol{X}_t}, X_t^i) - D_mF(m_t, \tilde{X}_t^i)|^2,\end{aligned}$$

where the difference between the drifts satisfies

$$\begin{aligned}&|D_mF(\mu_{\boldsymbol{X}_t}, X_t^i) - D_mF(m_t, \tilde{X}_t^i)| \\ &\quad\leqslant |D_mF(\mu_{\boldsymbol{X}_t}, X_t^i) - D_mF(\mu_{\tilde{\boldsymbol{X}}_t}, \tilde{X}_t^i)| + |D_mF(\mu_{\tilde{\boldsymbol{X}}_t}, \tilde{X}_t^i) - D_mF(m_t, \tilde{X}_t^i)| \\ &\quad\leqslant M_{mm}^F W_1(\mu_{\boldsymbol{X}_t}, \mu_{\tilde{\boldsymbol{X}}_t}) + M_{mx}^F |X_t^i - \tilde{X}_t^i| + M_{mm}^F W_1(\mu_{\tilde{\boldsymbol{X}}_t}, m_t).\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bigg(\sum_{i=1}^N |X_t^i - \tilde{X}_t^i|^2\bigg) \leqslant \big(1 + 3(M_{mx}^F)^2\big)\sum_{i=1}^N |X_t^i - \tilde{X}_t^i|^2 + 3N(M_{mm}^F)^2 W_2^2(\mu_{\boldsymbol{X}_t}, \mu_{\tilde{\boldsymbol{X}}_t}) \\ + 3N(M_{mm}^F)^2 W_2^2(\mu_{\tilde{\boldsymbol{X}}_t}, m_t). \quad (1.69)\end{aligned}$$

For the second term, we have

$$\mathbb{E}\big[W_2^2(\mu_{\boldsymbol{X}_t}, \mu_{\tilde{\boldsymbol{X}}_t})\big] \leqslant \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[|X_t^i - \tilde{X}_t^i|^2\big],$$

and for the last term, we have, by the result of Fournier and Guillin [93],

$$\begin{aligned}\mathbb{E}\big[W_2^2(\mu_{\tilde{\boldsymbol{X}}_t}, m_t)\big] &\leqslant C(d)\,\mathbb{E}\big[|X_t - \mathbb{E}\,X_t|^6\big]^{1/3}\delta_d(N) \\ &= C(d)\,\mathbb{E}\big[|X_t - \mathbb{E}\,X_t|^6\big]^{1/3} \times \begin{cases} N^{-1/2} & \text{if } d < 4, \\ N^{-1/2}\log(1+N) & \text{if } d = 4, \\ N^{-2/d} & \text{if } d > 4.\end{cases}\end{aligned}$$

Then, denoting $\tilde{X}_t = \tilde{X}_t^1$, we only need to control $\mathbb{E}\big[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^6\big]$. Observe that, by Itō's formula, we have

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^6] \\ &\quad= -6\,\mathbb{E}\big[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^4(\tilde{X}_t - \mathbb{E}\,\tilde{X}_t)\cdot\big(D_mF(m_t, \tilde{X}_t) - \mathbb{E}[D_mF(m_t, \tilde{X}_t)]\big)\big] \\ &\qquad + (6d + 24)\,\mathbb{E}\big[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^4\big].\end{aligned}$$

Then we have the following control of the growth, by using the elementary inequality $x^4 \leqslant \frac{2}{3}x^6 + \frac{1}{3}$ for $x \geqslant 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\big[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^6\big] \leqslant (6M_{mx}^F + 4d + 16)\,\mathbb{E}[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^6] + (2d + 8).$$

Thus, by Grönwall's lemma, we have

$$\mathbb{E}\big[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^6\big] \leqslant e^{(6M^F_{mx}+4d+16)t}\,\mathbb{E}\big[|\tilde{X}_0 - \mathbb{E}\,\tilde{X}_0|^6\big] + \frac{d+4}{3M^F_{mx}+2d+8}\big(e^{(6M^F_{mx}+4d+16)t}-1\big).$$

We take expectations on both side of the differential inequality (1.69) and obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\Bigg[\sum_{i=1}^N |X^i_t - \tilde{X}^i_t|^2\Bigg] \leqslant \big(1 + 3(M^F_{mx})^2 + 3(M^F_{mm})^2\big)\,\mathbb{E}\Bigg[\sum_{i=1}^N |X^i_t - \tilde{X}^i_t|^2\Bigg] + 3N(M^F_{mm})^2C(d)\delta_d(N)\,\mathbb{E}\big[|\tilde{X}_t - \mathbb{E}\,\tilde{X}_t|^6\big]^{1/3}.$$

We then use Grönwall's lemma to show (1.26).

As for the distance under relative entropy, by Girsanov's theorem we have

$$H(m^N_t|m^{\otimes N}_t) \leqslant H(m^N_0|m^{\otimes N}_0) + \frac{1}{4}\sum_{i=1}^N\int_0^t \mathbb{E}\big[|D_mF(\mu_{\boldsymbol{X}_s},X^i_s) - D_mF(m_s,X^i_s)|^2\big]\,\mathrm{d}s,$$

and we can control the last term by

$$\begin{aligned}|D_mF(\mu_{\boldsymbol{X}_s},X^i_s) - D_mF(m_s,X^i_s)| &\leqslant M^F_{mm}W_2(\mu_{\boldsymbol{X}_s},m_s)\\ &\leqslant M^F_{mm}\big(W_2(\mu_{\boldsymbol{X}_s},\mu_{\tilde{\boldsymbol{X}}_s}) + W_2(\mu_{\tilde{\boldsymbol{X}}_s},m_s)\big).\end{aligned}$$

So we can show (1.28) by using the same method as before.

Part 2: Long-time behavior. The triangle inequality for the L^2-Wasserstein distance gives us $W_2^2(m^N_t,m^{\otimes N}_t) \leqslant 2\big(W_2^2(m^N_t,m^{\otimes N}_\infty) + W_2^2(m^{\otimes N}_t,m^{\otimes N}_\infty)\big)$. By Talagrand's inequality (1.9) for $m^{\otimes N}_\infty$ we bound the Wasserstein distances by

$$\begin{aligned}\rho W_2^2(m^N_t,m^{\otimes N}_\infty) &\leqslant H(m^N_t|m^{\otimes N}_\infty) \leqslant \mathcal{F}^N(m^N_t) - N\mathcal{F}(m_\infty),\\ \rho W_2^2(m^{\otimes N}_t,m^{\otimes N}_\infty) &= NW_2^2(m_t,m_\infty) \leqslant NH(m^N_t|m_\infty) \leqslant N\big(\mathcal{F}(m_t) - \mathcal{F}(m_\infty)\big),\end{aligned}$$

where we applied Lemmas 1.40 and 1.52. We then apply Theorems 1.4 and 1.12 to obtain (1.25).

Now suppose additionally (1.6) and $h_0 = m_0/m_\infty \in L^{p_0}(m_\infty)$ for $p_0 > 1$. The relative entropy satisfies

$$\begin{aligned}H(m^N_t|m^{\otimes N}_t) &= \int \log\frac{m^N_t(\boldsymbol{x})}{m^{\otimes N}_t(\boldsymbol{x})}m^N_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= \int\left(\log\frac{m^N_t(\boldsymbol{x})}{m^{\otimes N}_\infty(\boldsymbol{x})} - \log\frac{m^{\otimes N}_t(\boldsymbol{x})}{m^{\otimes N}_\infty(\boldsymbol{x})}\right)m^N_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= H(m^N_t|m^{\otimes N}_\infty) - \sum_{i=1}^N\int\log\frac{m_t(x)}{m_\infty(x)}m^{N,i}_t(x)\,\mathrm{d}x,\end{aligned}$$

where $m^{N,i}_t$ is the i-th marginal of m^N_t. We then apply (1.54) in Lemma 1.49 to summands in the second term with $p=1$ to obtain

$$-\int\log\frac{m_t(x)}{m_\infty(x)}m^{N,i}_t(x)\,\mathrm{d}x \leqslant H(m^{N,i}_t|m_\infty) - \log\|h_t\|_{-1}.$$

So we have

$$-\sum_{i=1}^N \int \log \frac{m_t(x)}{m_\infty(x)} m_t^{N,i}(x)\,\mathrm{d}x \leqslant -N\log\|h_t\|_{-1} + \sum_{i=1}^N H(m_t^{N,i}|m_\infty)$$
$$\leqslant -N\log\|h_t\|_{-1} + H(m_t^N|m_\infty^{\otimes N}),$$

where we used the information inequality (1.56) in the last inequality. Therefore

$$H(m_t^N|m_t^{\otimes N}) \leqslant -N\log\|h_t\|_{-1} + 2H(m_t^N|m_\infty^{\otimes N}).$$

We conclude by applying the results of Theorems 1.9 and 1.12. □

Proof of Corollary 1.15. In the Wasserstein case, let C_4, C_5 be the constants in Theorem 1.14. We take $t_0 = \log N/(d\vee 4)C_4$. Then, for $t \leqslant t_0$, by using (1.26), we have

$$\frac{1}{N}W_2^2(m_t^N, m_t^{\otimes N}) \leqslant C_5(e^{C_4 t} - 1)\big(v_6(m_0)^{1/3} + 1\big)\delta_d(N)$$
$$\leqslant C_5(N^{1/(d\vee 4)} - 1)\big(v_6(m_0)^{1/3} + 1\big)\delta_d(N), \quad (1.70)$$

where $N^{1/(d\vee 4)}\delta_d(N) \leqslant N^{-1/(d\vee 4)}\log(1+N)$ for all d. For $t \geqslant t_0$, by using (1.25), we have

$$\frac{1}{N}W_2^2(m_t^N, m_t^{\otimes N}) \leqslant 2\big(\mathcal{F}(m_0) - \mathcal{F}(m_\infty)\big)N^{-4\rho/(d\vee 4)C_4}$$
$$+ \frac{2}{N}\big(\mathcal{F}^N(m_0^{\otimes N}) - N\mathcal{F}(m_\infty)\big)N^{-(4\rho' - C_1 N^{-1})/(d\vee 4)C_4}$$
$$+ \frac{2C_2}{4N\rho' - C_1}, \quad (1.71)$$

if $N > C_1/4\rho'$, where $\rho' \in (0,\rho)$ and C_1, C_2 are defined in Theorem 1.12. By expanding the functional F, we also have

$$F(\mu_{\boldsymbol{X}_0}) - F(m_0) = \int \frac{\delta F}{\delta m}(m_0, x)(\mu_{\boldsymbol{X}_0} - m_0)(\mathrm{d}x)$$
$$+ \int_0^1 \left(\frac{\delta F}{\delta m}\big((1-t)\mu_{\boldsymbol{X}_0} + tm_0, x\big) - \frac{\delta F}{\delta m}(m_0, x)\right)(\mu_{\boldsymbol{X}_0} - m_0)(\mathrm{d}x)dt$$

with

$$\mathbb{E}\left[\int \frac{\delta F}{\delta m}(m_0, x)(\mu_{\boldsymbol{X}_0} - m_0)(\mathrm{d}x)\right] = 0$$

and

$$\mathbb{E}\left[\int_0^1 \left(\frac{\delta F}{\delta m}\big((1-t)\mu_{\boldsymbol{X}_0} + tm_0, x\big) - \frac{\delta F}{\delta m}(m_0, x)\right)(\mu_{\boldsymbol{X}_0} - m_0)(\mathrm{d}x)\,\mathrm{d}t\right]$$
$$\leqslant \mathbb{E}\left[\int_0^1 \big\|D_m F\big((1-t)\mu_{\boldsymbol{X}_0} + tm_0, \cdot\big) - D_m F(m_0, \cdot)\big\|_\infty W_1(\mu_{\boldsymbol{X}_0}, m_0)\,\mathrm{d}t\right]$$
$$\leqslant \frac{M_{mm}^F}{2}\,\mathbb{E}\big[W_2^2(\mu_{\boldsymbol{X}_0}, m_0)\big] \leqslant M_{mm}^F \operatorname{Var} m_0.$$

Thus, we obtain

$$\begin{aligned}\mathcal{F}^N(m_0^{\otimes N}) = N\,\mathbb{E}\big[F(\mu_{\boldsymbol{X}_0})\big] + H(m_0^{\otimes N}) &\leqslant NF(m_0) + NM_{mm}^F \operatorname{Var} m_0 + NH(m_0)\\ &= N\mathcal{F}(m_0) + NM_{mm}^F \operatorname{Var} m_0. \qquad (1.72)\end{aligned}$$

Taking $\rho' = \rho/2$, we obtain the uniform-in-time Wasserstein bound (1.29) from (1.70) and (1.71).

Similarly, to control the relative entropy, we take $t_0' = \tau + \frac{\log N}{(d\vee 4)C_4}$, where τ is the constant in Theorem 1.14. So, for $t \leqslant t_0'$, by (1.28), we have

$$\frac{1}{N}H(m_t^N|m_t^{\otimes N}) \leqslant C_5(e^{C_4\tau}N^{1/(d\vee 4)} - 1)\big(v_6(m_0)^{1/3} + 1\big)\delta_d(N), \qquad (1.73)$$

and, for $t \geqslant t_0'$, by (1.27), we have

$$\begin{aligned}\frac{1}{N}H(m_t^N|m_t^{\otimes N}) &\leqslant C_3 e^{-4\rho'\tau}N^{-4\rho'/(d\vee 4)}\\ &\quad + \frac{2}{N}\big(\mathcal{F}^N(m_0^{\otimes N}) - N\mathcal{F}(m_\infty)\big)e^{-(4\rho' - C_1N^{-1})\tau}N^{-(4\rho'-C_1N^{-1})/(d\vee 4)C_4}\\ &\quad + \frac{2C_2}{4N\rho' - C_1}. \qquad (1.74)\end{aligned}$$

So, using again (1.72), we can combine (1.73) and (1.74) to obtain the uniform-in-time entropic bound (1.30). □

Chapter 2

Uniform-in-time propagation of chaos for kinetic mean field Langevin dynamics

Abstract. We study the kinetic mean field Langevin dynamics under the functional convexity assumption of the mean field energy functional. Using hypocoercivity, we first establish the exponential convergence of the mean field dynamics and then show the corresponding N-particle system converges exponentially in a rate uniform in N modulo a small error. Finally we study the short-time regularization effects of the dynamics and prove its uniform-in-time propagation of chaos property in both the Wasserstein and entropic sense. Our results can be applied to the training of two-layer neural networks with momentum and we include the numerical experiments.

Based on joint work with Fan Chen, Yiqing Lin and Zhenjie Ren.

2.1 Introduction

Training neural networks by momentum gradient descent has proven to be effective in various applications [213, 131, 204]. However, despite their excellent performance, the theoretical understanding of those algorithms remains elusive. Recently, extensive researches have been conducted to model the loss minimization of neural networks as a mean field optimization problem [163, 57, 203, 117], with most characterizing gradient descent algorithms as *overdamped mean field Langevin* (MFL) dynamics. In this paper, we will focus on *kinetic* dynamics instead, which corresponds to momentum gradient descent in the context of machine learning [185, 133]. Classical studies, such as [221, 157], have explored the exponential convergence of linear kinetic Langevin dynamics based on hypocoercivity and functional inequalities. The kinetic MFL dynamics is studied in [128] to model the momentum gradient descent for the training of neural networks and its convergence to the unique invariant measure is proven without a quantitative rate. The present work studies both the quantitative long-time behavior of the kinetic MFL dynamics and its uniform-in-time *propagation of chaos* (POC) property, under a *functional convexity* assumption, and we aim to provide a theoretical justification for the momentum algorithm's efficiency in practice.

2.1.1 Settings and main results

We give an informal preview of our settings and main results in this section. Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a *mean field functional* and denote by $D_m F : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ its *intrinsic derivative*. We aim to investigate the long-time behavior of the kinetic MFL defined by

$$\begin{aligned} dX_t &= V_t dt, \\ dV_t &= -V_t dt - D_m F\big(m_t^x, X_t\big) dt + \sqrt{2} dW_t, \qquad \text{where } m_t^x = \mathrm{Law}(X_t), \end{aligned}$$

and its associated N-particle system defined by

$$\begin{aligned} dX_t^i &= V_t^i dt \\ dV_t^i &= -V_t^i dt - D_m F\big(\mu_{\boldsymbol{X}_t}, X_t^i\big) dt + \sqrt{2} dW_t^i, \qquad \text{where } \mu_{\boldsymbol{X}_t} = \frac{1}{N} \sum_{j=1}^N \delta_{X_t^j}. \end{aligned}$$

Here W_t, $(W_t^i)_{i=1}^N$ are independent d-dimensional Brownian motions. Denote $m_t = \mathrm{Law}(X_t, V_t)$ and $m_t^N = \mathrm{Law}\big(X_t^1, \ldots, X_t^N, V_t^1, \ldots, V_t^N\big)$, and we suppose the initial conditions m_0 and m_0^N have finite second moments. We wish to show the convergence $m_t^N \to m_t^{\otimes N}$ when $N \to +\infty$ in a uniform-in-t way.

We assume

- the mean field functional F is convex in the functional sense;
- its intrinsic derivative $(m, x) \mapsto D_m F(m, x)$ is jointly Lipschitz with respect to the L^1-Wasserstein distance.
- for every measure $m \in \mathcal{P}_2(\mathbb{R}^d)$, the probability measure proportional to $\exp\big(-\frac{\delta F}{\delta m}(m, x)\big) dx$ satisfy a *logarithmic Sobolev inequality* (LSI) with a constant uniform in m.
- its second and third-order functional derivatives satisfy certain bounds.

Under these assumptions, we are able to obtain

- when $t \to +\infty$, the mean field flow m_t converges exponentially to the mean field invariant measure m_∞;
- when $t \to +\infty$, the N-particle flow m_t^N converges approximately to the N-tensorized mean field invariant measure $m_\infty^{\otimes N}$, with an exponential rate uniform in N;
- if $\frac{1}{N} W_2^2\big(m_0^N, m_0^{\otimes N}\big) \to 0$ when $N \to +\infty$, then $\sup_{t \geqslant 0} \frac{1}{N} W_2^2\big(m_t^N, m_t^{\otimes N}\big) \to 0$ and $\sup_{t \geqslant 1} \frac{1}{N} H\big(m_t^N \big| m_t^{\otimes N}\big) \to 0$ when $N \to +\infty$.

2.1.2 Related works

We give in this section a short review of the recent progresses in the long-time behavior and the uniform-in-time propagation of chaos property of McKean–Vlasov dynamics, with an emphasis on kinetic ones. We refer readers to [43, 44] for a more comprehensive review of propagation of chaos.

Coupling approaches. The coupling approach involves constructing a joint probability of the mean field and N-particle systems to allow comparisons between them. The *synchronous coupling* method is employed in [24] and the uniform-in-time POC is shown by assuming the strong monotonicity of the drift and the smallness of the mean field interaction. The strong monotonicity is then relaxed by the *reflection coupling* method in [84] and we refer readers to [206, 97, 128] for further developments. Let us remark that the synchronous coupling gives often sharp contraction rates under strong convexity assumptions, while the reflection coupling allows us to treat dynamics of more general type but gives far-from-sharp contraction rates.

Functional approaches. Another approach to uniform-in-time POC is the functional one, and this is also the major approach of this paper. In this situation in order to study the long-time behaviors and propagation of chaos properties, we construct appropriate (Lyapunov) functionals and investigate the change of their values along the dynamics. The relative entropy is used as the functional in [167] and its follow-up work [101] to study kinetic McKean–Vlasov dynamics with regular interactions. It is worth noting that the relative entropy approach has been successful in handling singular interactions, thanks to the groundbreaking work of Pierre-Emmanuel Jabin and Zhenfu Wang [124], and we refer the readers to [98, 59, 199] for recent developments. However, we are not aware of any works using the relative entropy functional (or its modifications) to study kinetic diffusions with singular interactions.

Comparison to Chapter 1. The present paper is a continuation of Chapter 1, where the overdamped version of mean field Langevin dynamics is studied, and they share a number of key features. We show the exponential convergence of the particle system using the same componentwise decomposition of Fisher information and the same componentwise log-Sobolev inequality. The uniform-in-time propagation of chaos property for both dynamics is then obtained by combining the exponential convergence of the mean field and particle flow. This paper is also different from Chapter 1 in a number of aspects. First, as the dynamics is generated by a hypoelliptic operator instead of an elliptic one, we use hypocoercivity to recover the exponential convergence. Second, since we are not able to show hypercontractivity of the kinetic dynamics (let alone reverse hypercontractivity), we prove the entropic propagation of chaos by studying its short-time regularization effects. In this way we no longer restrict the initial condition of the mean field dynamics, but as a trade-off we require a higher-order regularity in measure of the energy functional. Finally, following in a remark in Chapter 1, we use an approximation argument to remove the condition on the higher-order spatial derivatives in this work.

2.1.3 Main contributions

Hypocoercivity for mean field systems. We extend the studies of the linear Fokker–Planck equation in [221] to the dynamics with general (but always regular) mean field interactions. In particular, we do not suppose the interaction is in form of a two-body potential, which stands in contrast with [221, Theorem 56] and [167, 101]. Moreover, in hypocoercive computations, we find that the contributions from the mean field interaction can always be dominated by the "diagonal" terms in the Fisher information, already present in the case of linear dynamics. Hence using the convexity of energy, we are able to derive the hypocoercivity without restrictions on the size of the interaction. Furthermore, our assumptions imply a uniform-in-N bound on the operator norm of the second-order derivatives of the effective potential driving the N-particle system, and the entropic hypocoercivity is consequently uniform in N. This is different from the situation of L^2-hypocoercivity, where the condition given by Villani [221, (7.3)] gives dimension-dependent constants and therefore is unsuitable for studies of particle systems, as remarked in [99]. Finally, let us mention that we derive the entropic hypocoercivity under minimal regularity assumptions, made possible by our approximation argument (of functions *and* of mean field functionals) and the calculus in Wasserstein space developed in [4].

Regularization in short time. We obtain two short-time regularization results for the kinetic mean field dynamics. The first, from Wasserstein to entropy, is a consequence of the *logarithmic Harnack's inequality*, obtained by applying the *coupling by change of measure* method of Panpan Ren and Feng-Yu Wang in [193] to the mean field and N-particle diffusions. We remark also that very recently a similar inequality ([120, (3.13)]) is proved for the propagation of chaos of non-degenerate McKean–Vlasov diffusions. The second regularization, from entropy to Fisher information, is obtained by adapting Hérau's functional in [221] to our mean field setting and follows from the same hypocoercive computations as we prove the convergence of the mean field flow. We stress that although much stronger regularization phenomena are present, for example from measure initial values to L^p for every $p > 1$ and to H^k for every $k \geqslant 1$, our results have the advantage of growing at most linearly in dimension, making them suitable for studying the N-particle systems under the limit $N \to +\infty$.

Propagation of chaos. Finally, using the exponential convergence and the short-time regularizations, we derive the propagation of chaos for the kinetic MFL, i.e. bounds on the distances between the particle system and the mean field system. In particular, the initial value of the both systems can be arbitrary measures of finite second moments without any further regularity constraints. Moreover, the error terms do not have any dimension-dependence. It is noteworthy that our approach allows us to not rely on a uniform-in-time log-Sobolev inequality for the mean field flow, and also that the dynamics considered are realized on the whole space instead of the torus, standing in contrast with previous works, e.g. [98, 142, 70].

2.1.4 Notations

Let d be a positive integer and x, v be elements of $\mathbb{R}^d$. We denote the Euclidean norm of x and v by $|x|$ and $|v|$ respectively. The letter $z = (x, v)$ then denotes an element of $\mathbb{R}^{2d}$ with its Euclidean norm denoted by $|z|^2 = |x|^2 + |v|^2$. For a $d \times d$ real matrix M, we denote by $|M|_{\mathrm{op}}$ its operator norm with respect to the Euclidean metric of $\mathbb{R}^d$. Let $p \geqslant 1$. Define $\mathcal{P}_p(\mathbb{R}^d)$ to be the space of probabilities on $\mathbb{R}^d$ of finite p-moment, i.e. $\mathcal{P}_p(\mathbb{R}^d) = \{m \in \mathcal{P}(\mathbb{R}^d) : \int |x|^p m(dx) < +\infty\}$. We denote the L^p-Wasserstein metric by W_p and refer readers to [4, Chapter 7] for its definition and elementary properties.

Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a mean field functional. Denote by $\frac{\delta F}{\delta m} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ its linear functional derivative and by $D_m F = \nabla \frac{\delta F}{\delta m} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ its intrinsic derivative, if they exist. The definition of linear functional derivative on $\mathcal{P}_2(\mathbb{R}^d)$ can be found in [37, Definition 5.43].

Let X and Y be random variables. We denote the distribution of X by $\mathrm{Law}(X)$ and say $X \sim m$ if $m = \mathrm{Law}(X)$. We also say $X \stackrel{d}{=} Y$ if $\mathrm{Law}(X) = \mathrm{Law}(Y)$.

The set of couplings between probabilities μ and ν is denoted by $\Pi(\mu, \nu)$.

Let $N \geqslant 2$ be integer. The bold letters $\boldsymbol{x}_N = (x^1, \dots, x^N)$, $\boldsymbol{v}_N = (v^1, \dots, v^N)$ denote respectively N-tuples of elements in $\mathbb{R}^d$ and $\boldsymbol{z}_N = (z^1, \dots, z^N)$ an N-tuple of elements in $\mathbb{R}^{2d}$. We omit the subscript N when there are no ambiguities. Given $\boldsymbol{x}_N = (x^1, \dots, x^N) \in \mathbb{R}^{dN}$, we denote the corresponding empirical measure by

$$\mu_{\boldsymbol{x}_N} = \frac{1}{N} \sum_{i=1}^N \delta_{x^i}.$$

For $i = 1, \dots, N$, we define $-i = \{1, \dots, N\} \setminus \{i\}$, that is, the complementary index set, and we denote the empirical measures formed by the $N-1$ points $(x_j)_{j \neq i}$ by

$$\mu_{\boldsymbol{x}_N^{-i}} = \frac{1}{N-1} \sum_{j \neq i} \delta_{x^j}.$$

For an $\mathbb{R}^{dN}$-valued random variable $\boldsymbol{X}_N = (X^i)_{i=1}^N$, we can form the random empirical measures $\mu_{\boldsymbol{X}_N}, \mu_{\boldsymbol{X}_N^{-i}}$.

Let $I \subset \{1, \dots, N\}$ and $J = \{1, \dots, N\} \setminus I$ be the complementary index set. Let $\boldsymbol{Z}$ be an $\mathbb{R}^{2dN}$-valued random variable and and m^N be its distribution, belonging to $\mathcal{P}(\mathbb{R}^{2dN})$. We denote the marginal and the (regular) conditional distributions of m^N by

$$\begin{aligned} m^{N,I} &= \mathrm{Law}(Z^i)_{i \in I}, \\ m^{N,I|J}(\boldsymbol{z}^J) &= \mathrm{Law}\big((Z^i)_{i \in I} \big| Z_t^j = z^j,\ j \in J\big), \end{aligned}$$

where the latter is defined $m^{N,J}$-almost surely and $\boldsymbol{z}^J$ denotes the tuple $(z^j)_{j \in J}$. We identify i with the singleton $\{i\}$ when working with indices.

Whenever a measure $m \in \mathcal{P}(\mathbb{R}^d)$ has a density with respect to the d-dimensional Lebesgue measure, we denote its density function by m equally. The relative $H(\cdot|\cdot)$ between probabilities are always well defined and the absolute entropy $H(\cdot)$ is also well defined if the measure in the argument has finite second moment. If a measure $m \in \mathcal{P}(\mathbb{R}^d)$ has distributional derivatives Dm representable by a finite Borel measure and Dm is absolutely continuous with respect to m, we define its Fisher information by

$$I(m) = \int \left| \frac{Dm}{m} \right|^2 m,$$

where $\frac{Dm}{m}$ is the Radon–Nikodým derivative. Otherwise we set $I(m) = +\infty$. One can verify that $I(m)$ is finite only if $m \in W^{1,1}(\mathbb{R}^d)$[1], and in this case $I(m) = \int |\nabla m|^2/m$, where ∇m is the weak derivatives in $L^1(\mathbb{R}^d; \mathbb{R}^d)$. The Fisher information defined in this way corresponds to the functional considered in [3, (2.26)]. If m is a measure on $\mathbb{R}^d$ having finite Fisher information, and if γ is another measure

[1] We sketch the proof here. Suppose m has finite Fisher information. Set $m^n = m \star \rho^n$ for a mollifying sequence $(\rho^n)_{n \in \mathbb{N}}$. Then we have $\|m^n\|_{W^{1,1}} \leqslant C$ for all $n \in \mathbb{N}$. By Gagliardo–Nirenberg, m^n is uniformly bounded in L^p for some $p > 1$, so upon an extraction of subsequence, $(m^n)_{n \in \mathbb{N}}$ converges to some $m' \in L^p$ weakly. But $m^n \to m$ in $\mathcal{P}$. The two limits coincide, i.e. $m = m'$. Hence m has density with respect to the Lebesgue and so does Dm.

on $\mathbb{R}^d$ having weakly differentiable density with respect to the Lebesgue, we define the relative Fisher information by

$$I(m|\gamma) = \int \left| \frac{\nabla m}{m} - \frac{\nabla \gamma}{\gamma} \right|^2 m.$$

For nonnegative functions $f : \mathbb{R}^d \to [0, +\infty)$ we define its entropy by

$$\operatorname{Ent}_m f = \mathbb{E}_m[f \log f] - \mathbb{E}_m[f] \log \mathbb{E}_m[f],$$

which is well defined in $[0, +\infty]$ by Jensen's inequality.

Organization of paper. In Section 2.2, we introduce our assumptions, define the kinetic mean field Langevin and the particle system, and state our main results. We provide in Section 2.3 an exemplary dynamics modeling neural networks' training and present our numerical experiments. Moving on to the proofs, we first show in Section 2.4 the exponential convergence of the mean field and particle system dynamics. We then study in Section 2.5 finite-time propagation of chaos and regularizations of the kinetic MFL before combining all previous results and showing the propagation of chaos theorem in its full form. Finally, several technical results are proved in the appendices.

2.2 Assumptions and main results

Assumptions. Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a mean field functional. We suppose F is convex in the sense that for every $t \in [0, 1]$ and every m, $m' \in \mathcal{P}_2(\mathbb{R}^d)$,

$$F\big((1-t)m + tm'\big) \leqslant (1-t)F(m) + tF(m'). \tag{2.1}$$

Suppose also its intrinsic derivative $D_m F : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ exists and satisfies

$$\begin{aligned}&\forall x, x' \in \mathbb{R}^d,\ \forall m, m' \in \mathcal{P}_2(\mathbb{R}^d),\\ &\qquad |D_m F(m,x) - D_m F(m',x')| \leqslant M^F_{mm} W_1(m,m') + M^F_{mx}|x - x'|\end{aligned} \tag{2.2}$$

for some constants M^F_{mm}, $M^F_{mx} \geqslant 0$. For each $m \in \mathcal{P}_2(\mathbb{R}^d)$ we define a probability measure $\Pi^x(m)$ on $\mathbb{R}^d$ by $\Pi^x(m)(dx) \propto \exp\big(-\frac{\delta F}{\delta m}(m,x)\big)dx$ and suppose $\Pi^x(m)$ satisfies the *ρ^x-logarithmic Sobolev inequality* (LSI), uniformly in m, for some $\rho^x > 0$, that is, for every $m \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\forall f \in C^1_b(\mathbb{R}^d), \qquad \rho^x \operatorname{Ent}_{\Pi^x(m)}(f^2) \leqslant \mathbb{E}_{\Pi^x(m)}\big[|\nabla f|^2\big]. \tag{2.3}$$

Finally for some of the results we suppose additionally that F is third-order differentiable in measure with $\sup_{m \in \mathcal{P}_2(\mathbb{R}^d)} \sup_{x,x' \in \mathbb{R}^d} \big|D^2_m F(m,x,x')\big|_{\mathrm{op}} \leqslant M^F_{mm}$ and

$$\begin{aligned}&\forall m, m' \in \mathcal{P}_2(\mathbb{R}^d),\ \forall x \in \mathbb{R}^d,\\ &\left| \iint \left[\nabla_x \frac{\delta^3 F}{\delta m^3}(m', x, x', x') - \nabla_x \frac{\delta^3 F}{\delta m^3}(m', x, x', x'') \right] m(dx') m(dx'') \right|\\ &\qquad\qquad\qquad\qquad \leqslant M^F_{mmm}\end{aligned} \tag{2.4}$$

for some constant M^F_{mmm}.

Definition of $\hat{m}$ and functional inequalities. For each $m \in \mathcal{P}(\mathbb{R}^{2d})$, we define $\hat{m}$ to be the probability on $\mathbb{R}^{2d}$ satisfying

$$\hat{m}(dxdv) \propto \exp\Big(-\frac{\delta F}{\delta m}(m^x, x) - \frac{1}{2}|v|^2\Big) dxdv, \tag{2.5}$$

where m^x is the spatial marginal of m. Sometimes we will abuse the notation and define for a measure $m'^x \in \mathcal{P}_2(\mathbb{R}^d)$, the probability $\widehat{m'^x}(dxdv) \propto \exp\big(-\frac{\delta F}{\delta m}(m'^x, x) - \frac{1}{2}|v|^2\big)dxdv$. If F satisfies (2.3) with the LSI constant ρ^x, then setting

$$\rho = \rho^x \wedge \frac{1}{2}, \tag{2.6}$$

we have that the ρ-LSI holds for $\hat{m}$: for every $f \in C_b^1(\mathbb{R}^{2d})$,

$$\rho \operatorname{Ent}_{\hat{m}}(f^2) \leqslant \mathbb{E}_{\hat{m}}\big[|\nabla f|^2\big]. \tag{2.7}$$

As a consequence, we have the *Poincaré inequality*: for every $f \in \mathcal{C}_b^1(\mathbb{R}^{2d})$,

$$2\rho \operatorname{Var}_{\hat{m}}(f) \leqslant \mathbb{E}_{\hat{m}}\big[|\nabla f|^2\big]; \tag{2.8}$$

and *Talagrand's T_2 transport inequality*: for every $\mu \in \mathcal{P}_2(\mathbb{R}^{2d})$,

$$2\rho W_2^2(\mu, \hat{m}) \leqslant H(\mu|\hat{m}). \tag{2.9}$$

Mean field and particle system. We study the *mean field kinetic Langevin dynamics*, that is, the following McKean–Vlasov SDE

$$\begin{aligned} dX_t &= V_t dt, \\ dV_t &= -V_t dt - D_m F\big(m_t^x, X_t\big)dt + \sqrt{2} dW_t, \quad \text{where } m_t^x = \operatorname{Law}(X_t). \end{aligned} \tag{2.10}$$

Let $N \geqslant 2$. The corresponding N*-particle system* is defined by

$$\begin{aligned} dX_t^i &= V_t^i dt \\ dV_t^i &= -V_t^i dt - D_m F\big(\mu_{\boldsymbol{X}_t}, X_t^i\big)dt + \sqrt{2} dW_t^i, \quad \text{where } \mu_{\boldsymbol{X}_t} = \frac{1}{N}\sum_{j=1}^N \delta_{X_t^j}. \end{aligned} \tag{2.11}$$

Here W and W^i are standard Brownian motions in $\mathbb{R}^d$, and $(W^i)_{i=1}^N$ are independent. Their marginal distributions $m_t = \operatorname{Law}(X_t)$, $m_t^N = \operatorname{Law}(\boldsymbol{X}_t) = \operatorname{Law}(X_t^1, \ldots, X_t^N)$ solve respectively the Fokker–Planck equations:

$$\partial_t m = \Delta_v m + \nabla_v \cdot (mv) - v \cdot \nabla_x m + D_m F\big(m_t^x, x\big) \cdot \nabla_v m, \tag{2.12}$$

$$\partial_t m^N = \sum_{i=1}^N \Big(\Delta_{v^i} m^N + \nabla_{v^i} \cdot (m^N v^i) - v^i \cdot \nabla_{x^i} m^N + D_m F(\mu_{\boldsymbol{x}}, x^i) \cdot \nabla_{v^i} m^N\Big), \tag{2.13}$$

where on the second line $\mu_{\boldsymbol{x}} := \frac{1}{N}\sum_{i=1}^N \delta_{x^i}$. The mean field equation (2.12) is non-linear while the N-particle system equation (2.13) is linear. We will show in Lemma 2.12 the wellposedness of the mean field dynamics (2.12) with initial conditions of finite second moment.

Remark 2.1. We have fixed the volatility and the friction constants to simplify the computations. In order to apply our results to the diffusion process defined by

$$\begin{aligned} dX_t &= \alpha V_t dt, \\ dV_t &= -\gamma V_t dt - D_m F\big(\mathrm{Law}(X_t), X_t\big)dt + \sigma dW_t, \end{aligned} \tag{2.14}$$

with α, γ, $\sigma > 0$, we introduce the new variables:

$$x' = \frac{(2\gamma^3)^{1/2}}{\alpha\sigma}x, \quad v' = \frac{(2\gamma)^{1/2}}{\sigma}v, \quad t' = \gamma^{-1}t,$$

define m' to be the push-out of measure m under $x \mapsto x'$, and set

$$F'(m') = \left(\frac{2}{\gamma\sigma^2}\right)^{1/2} F(m).$$

Then the stochastic process $t' \mapsto (X'_{t'}, V'_{t'})$ satisfy

$$\begin{aligned} dX'_{t'} &= V'_{t'}dt', \\ dV'_{t'} &= -V'_{t'}dt' - D_m F'\big(\mathrm{Law}(X'_{t'}), X'_{t'}\big)dt + \sqrt{2}dW'_{t'}, \end{aligned}$$

where $W'_{t'} := \gamma^{1/2}W_t$ is a standard Brownian motion. In the same way we can treat the particle system defined by

$$\begin{aligned} dX^i_t &= \alpha V^i_t dt \\ dV^i_t &= -\gamma V^i_t dt - D_m F\big(\mu_{\boldsymbol{X}_t}, X^i_t\big)dt + \sigma dW^i_t, \quad \text{where } \mu_{\boldsymbol{X}_t} = \frac{1}{N}\sum_{j=1}^N \delta_{X^j_t}. \end{aligned} \tag{2.15}$$

Free energies and invariant measures. For measures $m \in \mathcal{P}_2(\mathbb{R}^{2d})$, $m^N \in \mathcal{P}_2(\mathbb{R}^{2dN})$, we introduce the mean field and N-particle free energies:

$$\mathcal{F}(m) = F(m^x) + \frac{1}{2}\int |v|^2 m(dxdv) + H(m), \tag{2.16}$$

$$\mathcal{F}^N(m^N) = \int \left(NF(\mu_{\boldsymbol{x}}) + \frac{1}{2}|\boldsymbol{v}|^2\right) m^N(d\boldsymbol{x}d\boldsymbol{v}) + H(m^N). \tag{2.17}$$

The functionals are well defined with values in $(-\infty, +\infty]$. We will also work with probability measures, $m_\infty \in \mathcal{P}_2(\mathbb{R}^{2d})$ and $m^N_\infty \in \mathcal{P}_2(\mathbb{R}^{2dN})$, satisfying

$$m_\infty(dxdv) \propto \exp\left(-\frac{\delta F}{\delta m}(m^x_\infty, x) - \frac{1}{2}|v|^2\right)dxdv, \tag{2.18}$$

$$m^N_\infty(d\boldsymbol{x}d\boldsymbol{v}) \propto \exp\left(-NF(\mu_{\boldsymbol{x}}) - \frac{1}{2}|\boldsymbol{v}|^2\right)d\boldsymbol{x}d\boldsymbol{v}, \tag{2.19}$$

and having finite exponential moments, that is to say, both the integrals

$$\int \exp\big(\alpha(|x| + |v|)\big)m_\infty(dxdv), \qquad \int \exp\big(\alpha(|\boldsymbol{x}| + |\boldsymbol{v}|)\big)m^N_\infty(d\boldsymbol{x}d\boldsymbol{v})$$

are finite for every $\alpha \geqslant 0$. We call m_∞, m^N_∞ *invariant* measures to the dynamics (2.12), (2.13) respectively. The existence and uniqueness of the invariant measures are guaranteed by our assumptions (2.1), (2.2) and (2.3), as will be stated in Lemma 2.8.

Main results. Recall that m_t and m_t^N are the respective marginal distributions of the mean field and the N-particle system (2.10), (2.11). We first prove the exponential entropic convergence result for the MFL dynamics (2.10).

Theorem 2.2 (Entropic convergence of MFL)**.** *Assume F satisfies (2.1), (2.2) and (2.3). If m_0 has finite second moment, finite entropy and finite Fisher information, then there exist constants*

$$C_0 = C_0\big(M_{mx}^F, M_{mm}^F\big), \qquad \kappa = \kappa\big(\rho^x, M_{mx}^F, M_{mm}^F\big)$$

such that for every $t \geqslant 0$,

$$\mathcal{F}(m_t) - \mathcal{F}(m_\infty) \leqslant \big(\mathcal{F}(m_0) - \mathcal{F}(m_\infty) + C_0 I(m_0|\hat{m}_0)\big)e^{-\kappa t}. \tag{2.20}$$

The proof of the theorem is postponed to Section 2.4.2. We note that the proof only relies on the W_2-Lipschitz continuity of $m \mapsto D_m F(m,x)$, contrary to the W_1 one stated in (2.2).

Our second major contribution is the uniform-in-N exponential entropic convergence of the particle systems.

Theorem 2.3 (Entropic convergence of particle systems)**.** *Assume F satisfies (2.1), (2.2) and (2.3). If m_0^N has finite second moment, finite entropy and finite Fisher information for some $N \geqslant 2$, then there exist constants*

$$C_0 = C_0\big(M_{mx}^F, M_{mm}^F\big), \quad C_1 = C_1\big(\rho^x, M_{mx}^F, M_{mm}^F\big), \quad \kappa = \kappa\big(\rho^x, M_{mx}^F, M_{mm}^F\big)$$

such that if $N > C_1/\kappa$, then for every $t \geqslant 0$,

$$\begin{aligned}\mathcal{F}^N\big(m_t^N\big) - N\mathcal{F}(m_\infty) \leqslant \Big(\mathcal{F}\big(m_0^N\big) - N\mathcal{F}(m_\infty) + C_0 I\big(m_0^N\big|m_\infty^N\big)\Big)e^{-(\kappa - C_1/N)t} \\ + \frac{C_1 d}{\kappa - C_1/N}.\end{aligned} \tag{2.21}$$

The proof of the theorem is postponed to Section 2.4.3.

Remark 2.4. Strictly speaking, the result (2.21) does not imply that the particle systems converge uniformly. We only show $\frac{1}{N}\mathcal{F}^N\big(m_t^N\big)$ approaches the mean field minimum $\mathcal{F}(m_\infty)$ uniformly quickly until they are $O(N^{-1})$-close to each other.

Remark 2.5. Theorems 2.2 and 2.3 state results concerning the convergence of the respective free energies, which we will also call "convergence of entropy" or "entropic convergence", since in both cases the differences of free energies are related to relative entropies, as shown in Lemmas 2.9 and 2.10.

We now present the main theorem, which establishes the uniform-in-time propagation of chaos in both the Wasserstein distance and the relative entropy. The results are direct consequences of the exponential convergence in Theorems 2.2 and 2.3 and the regularization phenomena to be studied in Section 2.5.

Theorem 2.6 (Wasserstein and entropic propagation of chaos)**.** *Assume F satisfies (2.1), (2.2), (2.3) and (2.4). If m_0 belongs to $\mathcal{P}_2(\mathbb{R}^d)$ and m_0^N belongs to $\mathcal{P}_2(\mathbb{R}^{dN})$ for some $N \geqslant 2$, then there exist constants $C_1 = C_1\big(M_{mx}^F, M_{mm}^F, M_{mmm}^F\big)$, $C_2 = C_2\big(\rho^x, M_{mx}^F, M_{mm}^F\big)$ and $\kappa = \kappa\big(\rho^x, M_{mx}^F, M_{mm}^F\big)$ such that if $N > C_2/\kappa$, then for* *every* $t > 0$,

$$\begin{aligned} W_2^2\big(m_t^N, m_t^{\otimes N}\big) &\\ \leqslant \min\bigg\{ & C_1 W_2^2\big(m_0^N, m_0^{\otimes N}\big)e^{C_1 t} + C_1(e^{C_1 t} - 1)(\operatorname{Var} m_0 + d), \\ & \frac{C_2 N}{(t\wedge 1)^6} W_2^2(m_0, m_\infty)e^{-\kappa t} + \frac{C_2}{(t\wedge 1)^6} W_2^2\big(m_0^N, m_\infty^{\otimes N}\big)e^{-(\kappa - C_2/N)t} \\ & + \frac{C_2 d}{\kappa - C_2/N}\bigg\}; \end{aligned} \tag{2.22}$$

moreover, for every t *and* s *such that* $s + 1 \geqslant t > s \geqslant 0$,

$$H\big(m_t^N \big| m_t^{\otimes N}\big) \leqslant \frac{C_1}{(t-s)^3} W_2^2(m_s^N, m_s^{\otimes N}) + C_1(e^{C_1(t-s)} - 1)(\operatorname{Var} m_s + d). \tag{2.23}$$

The proof of the theorem is postponed to Section 2.5.4.

Comments on the assumptions. Compared to Chapter 1, we have removed the technical assumption that $x \mapsto D_m F(m, x)$ has bounded higher-order derivatives by a mollifying procedure of the mean field functional. However, the spatial Lipschitz constant M_{mx}^F, appearing in the assumption (2.2), will contribute to the constants, especially the rate of convergence κ, in our theorems. Nevertheless, this behavior is expected for kinetic dynamics, as this dependency is already present for the linear Fokker–Planck dynamics in [221]. Finally, we introduce the new condition (2.4) on the second and third-order derivatives in measure of the mean field functional. The condition (2.4) is used to obtain $O(1)$ errors in the propagation of chaos bounds (2.22), (2.23) in Theorem 2.6, which are stronger than the dimension-dependent errors obtained from the method of Fournier and Guillin [93].

2.3 Application: training neural networks by momentum gradient descent

We have given in Section 1.3 several examples of mean field functionals satisfying conditions (2.1), (2.2) and (2.3) of our theorems, and the only additional condition that remains to verify is the bound on the higher-order measure derivative (2.4). In the following we will recall the mean field formulation of the loss of two-layer neural networks and its corresponding kinetic dynamics (see Examples 1.21 and 1.27), and verify that it satisfies indeed the additional assumption.

2.3.1 Mean field formulation of neural network

Recall that the structure of a two-layer neural network is determined by its *feature map*:

$$\mathbb{R}^d \ni z \mapsto \Phi(\theta; z) \coloneqq \ell(c)\varphi(a \cdot z + b) \in \mathbb{R}^{d'},$$

where $\theta \coloneqq (c, a, b) \in \mathbb{R}^{d'} \times \mathbb{R}^d \times \mathbb{R} =: S$ is the parameter of a single neuron, $\varphi : \mathbb{R} \to \mathbb{R}$ is a non-linear *activation function* satisfying the squashing condition (see (1.34)), and $\ell : \mathbb{R} \to [-L, L]$ is a *truncation function* with *threshold* $L \in (0, +\infty)$. Here the action of the truncation is tensorized: $\ell(c) = \ell(c^1, \ldots, c^{d'}) \coloneqq \big(\ell(c^1), \ldots, \ell(c^{d'})\big)$ for a d'-dimensional vector $c = (c^1, \dots, c^{d'})$. Then given N neurons with respective parameters $\theta^1, \dots, \theta^N$, the associated network's output reads

$$\mathbb{R}^d \ni z \mapsto \Phi^N(\theta^1, \dots, \theta^N; z) = \frac{1}{N}\sum_{i=1}^N \Phi(\theta^i; z) \in \mathbb{R}. \tag{2.24}$$

Here z should be considered as the input of the network, i.e. the *feature*, and the value $\Phi^N(\theta^1, \dots, \theta^N; z)$ should correspond to the *label*. We wish to find the optimal neuron parameters $(\theta_i)_{i=1}^N$ for a possibly unknown distribution of feature-label tuples $\mu \in \mathcal{P}(\mathbb{R}^{d+d'})$. In order to quantify the goodness of networks, we define the *loss*:

$$F^N_{\mathrm{NNet}}(\theta^1, \dots, \theta^N) = \frac{N}{2}\int |y - \Phi^N(\theta^1, \dots, \theta^N; z)|^2 \mu(dzdy). \tag{2.25}$$

It is proposed in [117] and Chapter 1 that instead of minimizing the original loss (2.25), we consider the mean field output function $\mathbb{E}^{\Theta\sim m}[\Phi(\Theta; \cdot)]$ and minimize the mean field loss

$$F^N_{\mathrm{NNet}}(m) = \int \big|y - \mathbb{E}^{\Theta\sim m}[\Phi(\Theta; z)]\big|^2 \mu(dzdy). \tag{2.26}$$

We also add a quadratic regularizer

$$F_{\mathrm{Ext}}(m) = \frac{\lambda}{2}\int |\theta|^2 m(d\theta)$$

with regularization parameter $\lambda > 0$. The final optimization problem then reads

$$\inf_{m\in\mathcal{P}_2(S)} F(m) \coloneqq F_{\mathrm{NNet}}(m) + F_{\mathrm{Ext}}(m). \tag{2.27}$$

Following the calculations in Chapter 1 we can show that if both the truncation and activation function are bounded and has bounded derivatives of up-to-second order, then the conditions (2.1), (2.2), (2.3) are verified. Finally, the third-order derivatives $\frac{\delta^3 F}{\delta m^3}$ is a constant thanks to the fact that the loss function is quadratic, and therefore the condition (2.4) is satisfied with $M^F_{mmm} = 0$.

Remark 2.7. Following Remark 1.28, we recognize that the SDE (2.10) describes the continuous version of the gradient descent algorithm with momentum. Among various momentum gradient descent methods commonly used to train neural networks, the most prevalent ones are RMSProp and Adam algorithms (see [112, 131]), where the momentum is accumulated and the step size is adapted along the dynamics. In [154, 208, 189] the authors studied the convergence of these momentum-based algorithms and compared them to algorithms without momentum based on optimization theory. We note that estimates of the discretization error and optimal parameters can also be found in these studies.

2.3.2 Numerical experiments

We present our numerical experiments in this section. Our experiments are based on the discretized version of a particle system dynamics (2.15). We first explain the optimization problem and the numerical algorithm, and then present our two experiments: the first investigates the convergence behavior as the number of particles tends to infinity, and the second compares the kinetic dynamics to the corresponding overdamped dynamics.

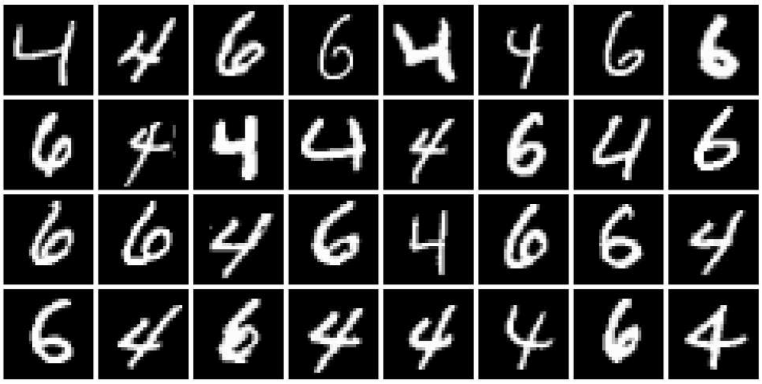


Figure 2.1: Randomly chosen handwritten digits “4” and “6” from the MNIST dataset.

Problem setup and momentum algorithm. We aim to solve a supervised learning problem:uklpoc- our goal is to classify the handwritten digits “4” and “6” by a two-layer neural network. We randomly choose $K = 10^4$ samples from the MNIST dataset [147] and denote by $(z_k)_{k=1}^K$ the figures in 28×28 pixel format, i.e. each z_k belongs to $\mathbb{R}^{28\times 28} = \mathbb{R}^{784}$, and by $(y_k)_{k=1}^K$ the *one-hot* vectors for the two classes of digits, i.e. if the k-th figure corresponds to the digit “4”, then $y_k = (1, 0)$, otherwise $y_k = (0, 1)$. See Figure 2.1 for random samples in the dataset. We choose N particles and use the sigmoid function as the activation, i.e. $\varphi(x) = 1/\bigl(1 + \exp(-x)\bigr)$. The truncation function is fixed by

$$\ell(x) = L \tanh(x/L) = L\,\frac{\exp(2x/L) - 1}{\exp(2x/L) + 1}$$

and its threshold equals L. The quadratic regularization parameter is denoted by λ. Following the arguments of Chapter 1 and the precedent section, all the conditions of our theorems (2.1), (2.2), (2.3), (2.4) are satisfied. In the beginning of training process, the neuron positions $(\Theta_0^i)_{i=1}^N = (C_0^{x,i}, A_0^{x,i}, B_0^{x,i})_{i=1}^N$ and momenta $(\Psi_0^i)_{i=1}^N = (C_0^{v,i}, A_0^{v,i}, B_0^{v,i})_{i=1}^N$ are sampled independently from a given initial distribution $m_0^x, m_0^v \in \mathcal{P}(\mathbb{R}^2 \times \mathbb{R}^{784} \times \mathbb{R})$. We update the parameters $(\Theta_0^i)_{i=1}^N$ and $(\Psi_0^i)_{i=1}^N$ following the discrete-time version of the underdamped Langevin SDE (2.14) with fixed set of parameters (α, γ, σ), that is, we calculate the neurons' evolution by Algorithm 2.

Convergence when $N \to +\infty$. To study the behavior of the momentum training dynamics when $N \to +\infty$ we conduct independent experiments with the an increasing number of particles: $N = 2^P$ for $P = 5, 6, \ldots, 10$ and repeat the experiment 10 times for each N. The hyperparameters for this experiment are listed in the second column of Table 2.1.

To quantify the convergence, we compute the two losses $\frac{1}{N} F_{\mathrm{NNet}}^N\bigl(\Theta_t^1, \ldots, \Theta_t^N\bigr)$ and $\frac{1}{N} F_{\mathrm{Kinet}}^N\bigl(\Psi_t^1, \ldots, \Psi_t^N\bigr)$, where $F_{\mathrm{Kinet}}^N(\Psi^1, \ldots, \Psi^N) := \frac{1}{2}\sum_{i=1}^N |\Psi^i|^2$. We then compute its average of the respective quantities over the 10 repeated runs. The evolutions of $\frac{1}{N} F_{\mathrm{NNet}}^N$ and $\frac{1}{N} F_{\mathrm{NNet}}^N + \frac{1}{N} F_{\mathrm{Kinet}}^N$ are plotted in Figures 2.2 and 2.3 respectively, and can be characterized by two distinct phases. In the first phase, both the quantities decrease and the second quantity decreases exponentially, for every N. We also find that in this phase the convergence rates are almost the same for different N and this is coherent with the behavior indicated by our theoretical upper bound (2.21). We also observe that $\frac{1}{N}F^N_{\text{NNet}}$ fluctuates in a stronger way than $\frac{1}{N}F^N_{\text{NNet}} + \frac{1}{N}F^N_{\text{Kinet}}$. In the second phase, both the values cease to decrease but the remnant values differ for different N.

To investigate the relationship between the remnant values in the second phase and the number of particles N, we compute the average value of $\frac{1}{N}F^N_{\text{NNet}} + \frac{1}{N}F^N_{\text{Kinet}}$ of the last 500 training epochs for each individual run and plot their values in Figure 2.4. Motivated by the upper bound (2.21) in Theorem 2.3, we fit the remnant values by $C' + \frac{C}{N}$ and find the values are well fitted by this curve.

Comparison to algorithm without momentum. We also investigate the difference between gradient descent algorithms with and without momentum by working on the same set of hyperparameters, listed in the last column of Table 2.1. It is found that the algorithm with momentum leads to much stronger fluctuations compared the algorithm without momentum (see Figure 2.5). Both algorithms cease to decrease after certain training epochs, but the momentum algorithm leads to better loss in the end. This may be explained by the fact that the presence of momentum helps the particles to escape local minima.

Hyperparameter	First Exp.'s Value	Second Exp.'s Value
N	$[128, 256, 512, 1024, 2048]$	256
Δt	0.02	0.01
T	300	500
m_0^x	$\mathcal{N}(0, 0.01)$	$\mathcal{N}(0, 0.01)$
m_0^v	$\mathcal{N}(0, 0.25)$	$\mathcal{N}(0, 0.01)$
L	500	500
λ	10^{-4}	10^{-3}
α	1	1
γ	0.1	0.1
σ	$0.01\sqrt{2}$	$0.01\sqrt{2}$

Table 2.1: Hyperparameters of neural networks' trainings.

2.4 Entropic convergence

2.4.1 Collection of known results

Before moving on to the proofs, we first state some elementary results without proofs. They are either immediate consequences of the corresponding ones in Chapter 1, or easy adaptations thereof.

Lemma 2.8 (Existence and uniqueness of invariant measures)**.** *If F satisfies* (2.1), (2.2) *and* (2.3)*, then there exist unique measures m_∞ and m^N_∞ satisfying* (2.18), (2.19) *respectively and they have finite exponential moments.*

Algorithm 2: Noised momentum gradient descent for training a two-layer neural network

---

Input: number of particles N, truncation L, data set $(z_k, y_k)_{k=1}^K$, noise σ, friction γ, l_2 regularization λ, initial distribution (m_0^x, m_0^v), time step Δt, time horizon T

Output: $(\Theta_T^i)_{i=1}^N$

generate i.i.d. $\Theta_0^i = \big(A_0^{x,i}, B_0^{x,i}, C_0^{x,i}\big) \sim m_0^x$ for $i = 1, \ldots, N$;

generate i.i.d. $\Psi_0^i = \big(A_0^{v,i}, B_0^{v,i}, C_0^{v,i}\big) \sim m_0^v$ for $i = 1, \ldots, N$;

for $t = 0, \Delta t, 2\Delta t, \ldots, T - \Delta t$ **do**

  generate i.i.d. $\mathcal{N}_t^i \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, N$;

  `// update particles according to discretized underdamped Langevin`

  for $i = 1, \ldots, N$ **do**

$$\Psi_{t+\Delta t}^i \leftarrow (1 - \gamma\Delta t)\Psi_t^i - D_m F_{\mathrm{NNet}}\big(\tfrac{1}{N}\textstyle\sum_{j=1}^N \delta_{\Theta_t^j}, \Theta_t^i\big)\Delta t - \lambda\Theta_t^i\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}_t^i;$$

$$\Theta_{t+\Delta t}^i \leftarrow \Theta_t^i + \Psi_{t+\Delta t}^i \Delta t;$$

  `// where` $D_m F_{\mathrm{NNet}}\big(\frac{1}{N}\sum_{j=1}^N \delta_{\Theta_t^j}, \Theta_t^i\big) = \frac{1}{K}\sum_{k=1}^K \big(y_k - \Psi^N(\Theta_t^1, \ldots, \Theta_t^N; z_k)\big)\frac{\partial\Psi}{\partial\theta}(\Theta_t^i; z_k)$

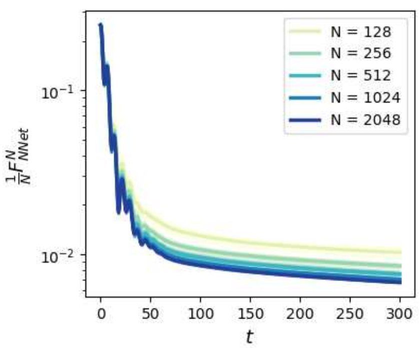


Figure 2.2: Individual (shadowed) and 10-averaged (bold) losses without kinetic energy versus time.

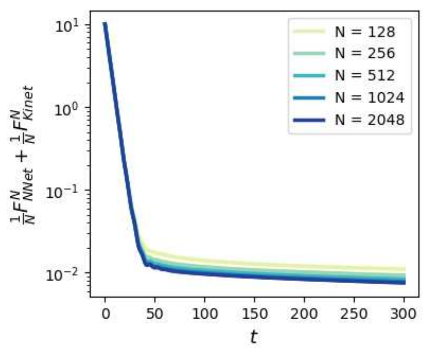


Figure 2.3: Individual (shadowed) and 10-averaged (bold) losses with kinetic energy versus time.

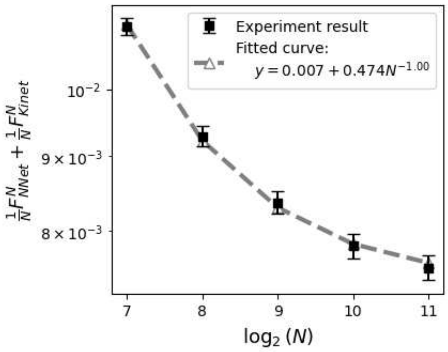


Figure 2.4: Average values of $\frac{1}{N}F_{\text{NNet}} + \frac{1}{N}F_{\text{Kinet}}$ over the last 500 epochs. The mean (black squares) and standard derivations (error bars) are calculated from the 10 independent runs. Dashed curve fits the data.

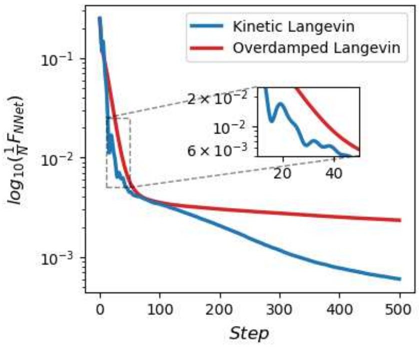


Figure 2.5: Target function $\frac{1}{N}F_{\text{NNet}}$ for underdamped Langevin (blue) and overdamped Langevin (red) versus time.

Lemma 2.9 (Mean field entropy sandwich)**.** *Assume F satisfies (2.1), (2.2) and (2.3). Then for every $m \in \mathcal{P}_2(\mathbb{R}^{2d})$, we have*

$$\begin{aligned} H(m|m_\infty) \leqslant \mathcal{F}(m) - \mathcal{F}(m_\infty) \leqslant H(m|\hat{m}) \\ \leqslant \left(1 + \frac{M^F_{mm}}{\rho} + \frac{(M^F_{mm})^2}{2\rho^2}\right) H(m|m_\infty), \end{aligned} \tag{2.28}$$

where ρ is defined by (2.6). Here, the leftmost inequality holds even without the uniform LSI condition (2.3), once there exists a measure m_∞ satisfying (2.18) and having finite exponential moments.

Lemma 2.10 (Particle system's entropy inequality)**.** *Assume that F satisfies (2.1) and that there exists a measure $m_\infty \in \mathcal{P}_2(\mathbb{R}^{2d})$ verifying (2.18). Then for all $m^N \in \mathcal{P}_2(\mathbb{R}^{dN})$ of finite entropy, we have*

$$H(m^N|m_\infty^{\otimes N}) \leqslant \mathcal{F}^N(m^N) - N\mathcal{F}(m_\infty). \tag{2.29}$$

Lemma 2.11 (Information inequalities)**.** *Let $X_1, \ldots, X_N$ be measurable spaces, μ be a probability on the product space $X = X_1 \times \cdots \times X_N$ and $\nu = \nu^1 \otimes \cdots \otimes \nu^N$ be a σ-finite measure. Then*

$$\sum_{i=1}^N H(\mu^i|\nu^i) \leqslant H(\mu|\nu) \leqslant \sum_{i=1}^N \int H\big(\mu^{i|-i}(\cdot|\boldsymbol{x}^{-i})\big|\nu^i\big)\mu^{-i}(d\boldsymbol{x}^{-i}). \tag{2.30}$$

Here we set the rightmost term to $+\infty$ if the conditional distribution $\mu^{i|-i}$ does not exist μ^{-i}-a.e.

2.4.2 Mean field system

In this section we study the mean field system described by the Fokker–Planck equation (2.12) and the SDE (2.10). Our aim is to prove Theorem 2.2. To this end, we first show its wellposedness and regularity.

Lemma 2.12. *Suppose F satisfies (2.2). Then for every initial value m_0 of finite second moment, the equation (2.12) admits a unique solution in $C\big([0,\infty);\mathcal{P}_2(\mathbb{R}^d)\big)$. Moreover, for every $t > 0$, the measure m_t is absolutely continuous with respect to the Lebesgue measure.*

Proof. Since the drift $D_m F(\cdot,\cdot)$ of the SDE system (2.10) is jointly Lipschitz in measure and in space by our condition (2.2), the existence and uniqueness of the solution is standard.

To show the existence of density we recall Kolmogorov's fundamental solution

$$\rho_t(x,v;x',v') \coloneqq \left(\frac{\sqrt{3}}{2\pi t^2}\right)^d \exp\biggl(-\frac{3|x-(x'+tv')|^2}{t^3} \\ + \frac{3\big(x-(x'+tv')\big)\cdot(v-v')}{t^2} - \frac{|v-v'|^2}{t}\biggr)$$

associated to the differential operator $\partial_t - \Delta_v + v\cdot\nabla_x$. Then the Duhamel's formula holds in the sense of distributions:

$$m_t = \int \rho_t(\cdot;z')m_0(dz') \\ + \int_0^t \iint \rho_s(\cdot;x',v')\nabla_{v'}\cdot\Big(m_{t-s}(dx'dv')\big(v'+D_mF(m^x_{t-s},x')\big)\Big)ds. \tag{2.31}$$

Since the first moment of m_t is bounded, that is, for every $T > 0$, $\sup_{t\in[0,T]}\int(|v| + |x|)m_t(dxdv) < +\infty$, we can integrate by parts in the second term of (2.31) and obtain

$$\|m_t\|_{L^1} \leqslant 1 + C\int_0^t \sup_{x',v'}\|\nabla_{v'}\rho_s(\cdot;x',v')\|_{L^1}ds.$$

By explicit computations we have $\sup_{x',v'}\|\nabla_{v'}\rho_s(\cdot;x',v')\|_{L^1} = O(s^{-1/2})$, from which the existence of the density follows. □

We now introduce a technical condition on the mean field functional: the mapping $x \mapsto D_mF(m,x)$ is fourth-order differentiable with derivatives continuous in measure and in space, and satisfying

$$\sup_{m\in\mathcal{P}_2(\mathbb{R}^d)}\sup_{x\in\mathbb{R}^d}\big|\nabla^k D_mF(m,x)\big| < +\infty, \qquad k = 2,\,3,\,4. \tag{2.32}$$

This condition will be used to derive some intermediate results in the following studies of the mean field dynamics.

Definition 2.13 (Standard algebra)**.** We define the *standard algebra* $\mathcal{A}_+$ to be the set of C^4 functions $h : \mathbb{R}^{2d} \to (0,\infty)$ for which there exists a constant C such that

$$|\log h(x,v)| \leqslant C(1+|x|+|v|) \quad \text{and} \quad \sum_{k=1}^4\big|\nabla^k h(x,v)\big| \leqslant \exp\big(C(1+|x|+|v|)\big)$$

holds for every $(x,v)\in\mathbb{R}^{2d}$. For a collection of functions $(h_\iota)_{\iota\in I}$ we say $h_\iota \in \mathcal{A}_+$ *uniformly* for $\iota \in I$ or $(h_\iota)_{\iota\in I} \subset \mathcal{A}_+$ *uniformly*, if there exists a constant C such that the previous bounds holds for every h_ι, $\iota \in I$.

Proposition 2.14 (Density of $\mathcal{A}_+$)**.** *Assume F satisfies* (2.2) *and* (2.32)*, and there exists a measure m_∞ satisfying* (2.18) *and having finite exponential moments. Then for every $m \in \mathcal{P}_2(\mathbb{R}^d)$ with finite entropy and finite Fisher information, there exists a sequence of measures $(m_n)_{n\in\mathbb{N}}$ such that $m_n/m_\infty \in \mathcal{A}_+$ and*

$$W_2(m_n, m) \to 0, \quad H(m_n) \to H(m), \quad I(m_n) \to I(m)$$

when $n \to +\infty$.

Proof. Let ε be arbitrary positive real. Put $h = m/m_\infty$. Define $h'_n = (h \wedge n) \vee \frac{1}{n}$ and the associated probability measure $m'_n = h'_n m_\infty / \int h'_n m_\infty$. Let $N \in \mathbb{N}$ be big enough so that $\int h'_N m_\infty > 0$. Note that

$$\sup_{n\geqslant N} |x|^2 m'_n(x) \leqslant \frac{|x|^2 m}{\int h'_N m_\infty},$$

that is to say, the second moments of $(m'_n)_{n\in\mathbb{N}}$ are uniformly bounded. Together with the fact that the density of m'_n converges to that of m pointwise, we have $m'_n \to m$ in $\mathcal{P}_2$. By the dominated convergence theorem, the sequence of measures m'_n satisfies

$$H(m'_n) = \frac{\int \log(h'_n m_\infty) h'_n m_\infty}{\int h'_n m_\infty} - \log \int h'_n m_\infty \to \int m \log m, \quad \text{when } n \to +\infty.$$

Moreover, we have the convergence of Fisher information as

$$\int \frac{|\nabla(h'_n m_\infty)|^2}{h'_n m_\infty} = \int \biggl[\frac{|\nabla m_\infty|^2 h'_n}{m_\infty} + \biggl(2\frac{\nabla h \cdot \nabla m_\infty}{h m_\infty} + \frac{|\nabla h|^2 m_\infty}{h} \biggr) \mathbb{1}_{1/n\leqslant h\leqslant n} \biggr]$$

converges to $I(m)$ when $n \to +\infty$, where we used the fact that the weak derivatives satisfy $\nabla h'_n = \nabla h \mathbb{1}_{1/n\leqslant h\leqslant n}$. Hence we may choose $n_0 \in \mathbb{N}$ such that

$$W_2(m_{n_0}, m) + \bigl|H(m_{n_0}) - H(m)\bigr| + |I(m_{n_0}) - I(m)| \leqslant \frac{\varepsilon}{2}.$$

Now set $m''_n = m'_{n_0} \star \eta_n$, where $(\eta_n)_{n\in\mathbb{N}}$ is a sequence of smooth mollifiers supported in the unit ball. We have $m''_n \to m'_{n_0}$ in $\mathcal{P}_2$. By the convexity of entropy and Fisher information we have $H(m''_n) \leqslant H\bigl(m'_{n_0}\bigr)$ and $I(m''_n) \leqslant I\bigl(m'_{n_0}\bigr)$, and by the lower semicontinuities in Lemma B.1, we have $\liminf_{n\to+\infty} H(m''_n) \geqslant H\bigl(m'_{n_0}\bigr)$ and $\liminf_{n\to+\infty} I(m''_n) \geqslant I\bigl(m'_{n_0}\bigr)$. Hence,

$$W_2\bigl(m''_n, m'_{n_0}\bigr) + \bigl|H\bigl(m''_n\bigr) - H\bigl(m'_{n_0}\bigr)\bigr| + \bigl|I\bigl(m''_n\bigr) - I\bigl(m'_{n_0}\bigr)\bigr| \to 0$$

when $n \to +\infty$. So we pick another $n_1 \in \mathbb{N}$ such that $W_2\bigl(m''_{n_1}, m'_{n_0}\bigr) + \bigl|H\bigl(m''_{n_1}\bigr) - H\bigl(m'_{n_0}\bigr)\bigr| + \bigl|I\bigl(m''_{n_1}\bigr) - I\bigl(m'_{n_0}\bigr)\bigr| < \varepsilon/2$.

It remains to verify that m''_{n_1}/m_∞ belongs to $\mathcal{A}_+$. By the definition we have

$$\frac{m''_{n_1}}{m_\infty} = \frac{(h'' m_\infty) \star \rho_{n_1}}{m_\infty}$$

for some h'' with $0 < \inf h'' \leqslant \sup h'' < +\infty$. Hence for every $z \in \mathbb{R}^{2d}$,

$$\inf h'' \frac{\inf_{B(z,1)} m_\infty}{m_\infty(z)} \leqslant \frac{m''_{n_1}(z)}{m_\infty(z)} \leqslant \sup h'' \frac{\sup_{B(z,1)} m_\infty}{m_\infty(z)}.$$

On the other hand, the gradient of m_∞ satisfies $|\nabla \log m_\infty(z)| \leqslant |D_m F(m_\infty, x)| + |v|$ for every $z = (x,v) \in \mathbb{R}^{2d}$. In particular, we have

$$\exp\bigl(-C(1+|x|+|v|)\bigr) \leqslant \frac{\inf_{B(z,1)} m_\infty}{m_\infty(z)} \leqslant \frac{\sup_{B(z,1)} m_\infty}{m_\infty(z)} \leqslant \exp\bigl(C(1+|x|+|v|)\bigr),$$

for some constant C. Therefore, m''_{n_1}/m_∞ verifies the first condition of $\mathcal{A}_+$.

Now verify the conditions on the derivatives. The derivatives read

$$\nabla^k \left(\frac{m''_{n_1}}{m_\infty}\right) = \sum_{j=0}^{k} \binom{k}{j} \frac{(h'' m_\infty) \star \nabla^j \rho_{n_1}}{m_\infty} \cdot m_\infty \nabla^{k-j}\bigl(m_\infty^{-1}\bigr).$$

For each term in the sum, we can bound its first part by

$$\left|\frac{(h'' m_\infty) \star \nabla^j \rho_{n_1}(z)}{m_\infty(z)}\right| \leqslant \exp\bigl(C(1+|z|)\bigr),$$

using the same method that we used to verify the first condition of $\mathcal{A}_+$. Moreover, since our assumptions (2.2), (2.32) imply

$$|\nabla \log m_\infty(z)| \leqslant C\bigl(1+|z|\bigr) \qquad \text{and} \qquad \bigl|\nabla^k \log m_\infty(z)\bigr| \leqslant C \quad \text{for } k = 2,\, 3,\, 4,$$

the second part of each term of the sum, $m_\infty \nabla^{k-j}\bigl(m_\infty^{-1}\bigr)$, is of polynomial growth. The proof is then complete. □

Then we show the stability of the set $\mathcal{A}_+$ under the mean field flow. This property will be used to justify the computations in the proof of Theorem 2.2, as is usual in the analysis of PDE.

Proposition 2.15 (Stability of $\mathcal{A}_+$ under flow)**.** *Assume that F satisfies (2.2) and (2.32), and that there exists a measure m_∞ satisfying (2.18) and having finite exponential moments. Let $(m_t)_{t\in[0,T]} \in C([0,T];\mathcal{P}_2(\mathbb{R}^d))$ be a solution in the sense of distributions to the mean field Fokker–Planck equation (2.12). If $m_0/m_\infty \in \mathcal{A}_+$, then $m_t/m_\infty \in \mathcal{A}_+$ uniformly for $t \in [0,T]$. In particular, m_t is a classical solution to the Fokker–Planck equation.*

Proof. In the following C will denote a constant depending on M^F_{mx}, M^F_{mm}, the initial value $h_0 = h(0,\cdot) \coloneqq m_0/m_\infty$, the time interval T and the bounds on the higher-order derivatives $\max_{k=2,3,4} \sup_{m,x} \bigl|\nabla^k D_m F(m,x)\bigr|$, and it may change from line to line. For a given quantity Q, we denote by C_Q a constant depending additionally on Q.

Denote $b_t(x) = -D_m F(m_t, x)$ and $b_\infty(x) = -D_m F(m_\infty, x)$. We also define $h_t(x) = m_t(x)/m_\infty(x)$. The relative density solves

$$\partial_t h = \Delta_v h - v \cdot \nabla_v h - v \cdot \nabla_x h - b_t \cdot \nabla_v h + (b_t - b_\infty) \cdot v h. \tag{2.33}$$

Fix $t \in [0,T]$. We construct for every $z = (x,v) \in \mathbb{R}^{2d}$, the stochastic process $Z^{t,z}_s = (X^{t,z}, V^{t,z})$, solving

$$\begin{aligned} dX^{t,z}_s &= -V^{t,z}_s ds, \\ dV^{t,z}_s &= -V^{t,z}_s ds - b_{t-s}\bigl(X^{t,z}_s\bigr) ds + \sqrt{2} dW_s \end{aligned}$$

for $s \in [0,t]$ with the initial value $X_0^{t,z} = x, V_0^{t,z} = v$ and the *same* Brownian motion $(W_s)_{s\in[0,t]}$.

Regularity of $Z_s^{t,z}$. Set $M^{t,z} = \sup_{s\in[0,t]}\bigl|Z_s^{t,z}\bigr|$. By Itō's formula and Doob's maximal inequality, the processes satisfy for every $\alpha \geqslant 0$,

$$\mathbb{E}\bigl[\exp(\alpha M^{t,z})\bigr] \leqslant \exp\bigl(C_\alpha(1+|z|)\bigr). \tag{2.34}$$

Thanks to the assumption on the uniform boundedness of the higher-order derivatives (2.32), the mapping $z \mapsto Z_s^{t,z}$ is C^4 and the partial derivatives solve the Cauchy–Lipschitz SDEs for $k = 1, 2, 3, 4$:

$$\begin{aligned} d\nabla^k X_s^{t,z} &= -\nabla^k V_s^{t,z} ds, \\ d\nabla^k V_s^{t,z} &= -\nabla^k V_s^{t,z} ds - \sum_{j=1}^k \nabla^j b_{t-s}\bigl(X_s^{t,z}\bigr) B_{k,j}\bigl(\nabla X_s^{t,z}, \dots, \nabla^{k-j+1} X_s^{t,z}\bigr) ds, \end{aligned}$$

where $B_{k,j}$ is a $k-j+1$-variate polynomial and in particular $B_{k,1}(x_1,\dots,x_k) = x_k$. The initial values of the SDEs read

$$\nabla Z_0^{t,z} = \mathrm{Id} \qquad \text{and} \qquad \nabla^k Z_0^{t,z} = 0 \quad \text{for } k = 2,\ 3,\ 4.$$

By induction we can obtain the almost sure bound

$$\max_{k=1,2,3,4} \sup_{s\in[0,t]} \bigl|\nabla^k Z_s^{t,z}\bigr| \leqslant C. \tag{2.35}$$

Regularity of h by Feynman–Kac. Denote $g(t,z) = g(t,x,v) = \bigl(b_t(x) - b_\infty(x)\bigr)\cdot v$. It satisfies

$$\bigl|g(t,z)\bigr| \leqslant M_{mm}^F W_2(m_t, m_\infty)|v| \leqslant M_{mm}^F \sup_{t\in[0,T]} W_2(m_t, m_\infty)|v| = C|v|$$

and also $\bigl|\nabla^k g(t,z)\bigr| \leqslant C\bigl(1+|z|\bigr)$ for $k = 1, 2, 3, 4$. The Feynman–Kac formula for the parabolic equation (2.33) reads

$$h(t,z) = \mathbb{E}\biggl[\exp\biggl(\int_0^t g\bigl(t-s, Z_s^{t,z}\bigr) ds\biggr) h\bigl(0, Z_t^{t,z}\bigr)\biggr]. \tag{2.36}$$

Using the method in the proof of Proposition 1.44, we can prove

$$|\log h(t,z)| \leqslant C(1+|z|).$$

Moreover, thanks to the estimates (2.35), we can apply the dominated convergence theorem to the Feynman–Kac formula (2.36) and obtain that $z \mapsto h(t,z)$ belongs to C^4 with partial derivatives

$$\begin{aligned} \nabla^k h(t,z) = \sum_{j=0}^k \mathbb{E}\biggl[\exp\biggl(\int_0^t g\bigl(t-s, Z_s^{t,z}\bigr) ds\biggr) P_j\biggl(\int_0^t \nabla_z g\bigl(t-s, Z_s^{t,z}\bigr) ds, \dots, \\ \int_0^t \nabla_z^j g\bigl(t-s, Z_s^{t,z}\bigr) ds\biggr) \nabla_z^{k-j} h\bigl(0, Z_t^{t,z}\bigr)\biggr], \end{aligned}$$

where P_j is a j-variate polynomial. Note that

$$\nabla_z^k f(Z_s^{t,z}) = \sum_{\ell=1}^{k} \nabla^\ell f\bigl(Z_s^{t,z}\bigr) B_{k,\ell}\bigl(\nabla Z_s^{t,z}, \dots, \nabla^{k-\ell+1} Z_s^{t,z}\bigr)$$

holds for $f = g(t-s,\cdot), s \in [0,t]$ and for $f = h(0,\cdot)$. We apply the bounds on $\bigl|\nabla^k g\bigr|, \bigl|\nabla^k h\bigr|$ for $k = 0, 1, 2, 3, 4$ and the exponential moment bound (2.34) to obtain that $\bigl|\nabla^k h(t,z)\bigr| \leqslant \exp\bigl(C(1+|z|)\bigr)$ for $k = 1, 2, 3, 4$. Finally, the derivatives $\nabla h, \nabla^2 h$ exist and one can show that they are continuous in time by differentiating (2.33) twice in space. So again by the equation (2.33) we have $\partial_t h$ is continuous and therefore exists classically. Thus m_t is a classical solution to the Fokker–Planck equation (2.12). □

Remark 2.16. The polynomials appearing in the previous proof belong to the *non-commutative* free algebras over $\mathbb{R}$ of respective number of indeterminates instead of the usual polynomial rings, as the tensor product is not commutative.

After the technical preparations we prove Theorem 2.2.

Proof of Theorem 2.2. The proof consists of several steps.

Step 1: Preparations. Suppose first that the mean field functional F satisfies additionally (2.32) and the initial value of the dynamics is such that m_0/m_∞ belongs to $\mathcal{A}_+$, which is the standard algebra defined in Definition 2.13. According to Proposition 2.15, the measure m_t belongs to $\mathcal{A}_+$ uniformly in t, for every $T > 0$. Since we have that $z \mapsto \hat{m}_t(z)/m_\infty(z)$ is C^4 with

$$\sup_{z\in\mathbb{R}^{2d}} \left|\nabla \log \frac{\hat{m}_t}{m_\infty}(z)\right| \leqslant M_{mm}^F W_2(m_t, m_\infty)$$

and

$$\max_{k=2,3,4} \sup_{z\in\mathbb{R}^{2d}} \left|\nabla^k \log \frac{\hat{m}_t}{m_\infty}(z)\right| \leqslant M,$$

for some constant M, the alternative relative density $\eta_t(z) \coloneqq m_t(z)/\hat{m}_t(z)$ is C^4 in z and there exists a constant M_T such that

$$\eta_t(z) + \frac{1}{\eta_t(z)} + \sum_{k=1}^{4} \bigl|\nabla^k \eta_t(z)\bigr| \leqslant \exp\bigl(M_T(1+|z|)\bigr) \tag{2.37}$$

for every $(t,z) \in [0,T] \times \mathbb{R}^{2d}$. The constant M_T may change from line to line in the following.

In the following we will adopt the abstract notations introduced by Villani in his seminal work on the hypocoercivity [221]. Define $\mathcal{H}_t = L^2(\hat{m}_t)$, $A_t = \nabla_v$ and $B_t = v \cdot \nabla_x - D_m F(m_t,x) \cdot \nabla_v$. The adjoint of A_t in $\mathcal{H}_t$ is therefore $A_t^* = -\nabla_v + v$, while B_t is antisymmetric: $B_t^* = -B_t$. Define the commutator $C_t = [A_t, B_t] = A_t B_t - B_t A_t = \nabla_x$. Finally define $L_t = A_t^* A_t + B_t$ and $u_t = \log \eta_t$. The Fokker–Planck equation (2.12) now reads

$$\frac{\partial_t m_t}{\hat{m}_t} = -L_t \eta_t = -(A_t^* A_t + B_t)\eta_t. \tag{2.38}$$

Step 2: Adding anisotropic Fisher. Let a, b, c be positive reals to be determined. We define the hypocoercive Lyapunov functional

$$\mathcal{E}(m) = \mathcal{F}(m) + a\int\left|\nabla_v \log\frac{m}{\hat{m}}(z)\right|^2 m(dz) \\ + 2b\int \nabla_v \log\frac{m}{\hat{m}}(z)\cdot\nabla_x \log\frac{m}{\hat{m}}(z)m(dz) + c\int\left|\nabla_x \log\frac{m}{\hat{m}}(z)\right|^2 m(dz), \quad (2.39)$$

where $\mathcal{F}(m) = F(m) + \frac{1}{2}\int|v|^2 m + H(m)$ is the free energy. We also denote the sum of the last three terms in (2.39) by $I_{a,b,c}(m_t|\hat{m}_t)$, so that

$$\mathcal{E}(m) = \mathcal{F}(m) + I_{a,b,c}(m_t|\hat{m}_t).$$

Thanks to Proposition 2.15 and in particular the bound (2.37), we can show that the quantity $\mathcal{E}(m_t)$ is well defined for every $t \geqslant 0$ and is continuous in t. We will show in the following that $t \mapsto \mathcal{E}(m_t)$ is in fact absolutely continuous and calculate its almost everywhere derivative. To this end, for every $t > 0$ and every $h \geqslant -t$, we define

$$\begin{aligned}\mathcal{E}(m_{t+h}) - \mathcal{E}(m_t) &= \big(\mathcal{F}(m_{t+h}) - \mathcal{F}(m_t)\big) \\ &\quad + \big(I_{a,b,c}(m_{t+h}|\hat{m}_{t+h}) - I_{a,b,c}(m_t|\hat{m}_{t+h})\big) \\ &\quad + \big(I_{a,b,c}(m_t|\hat{m}_{t+h}) - I_{a,b,c}(m_t|\hat{m}_t)\big) \\ &=: \Delta_1 + \Delta_2 + \Delta_3.\end{aligned}$$

Step 3: Contributions from Δ_1 and Δ_2. We first calculate the contributions from Δ_1. Using the Fokker–Planck equation (2.12) and the bounds (2.37), one has $|\Delta_1| \leqslant M_T h$ for every t, h such that t and $t+h$ belong to $[0, T]$; moreover, by the dominated convergence theorem one has for almost every $t > 0$,

$$\lim_{h\to 0}\frac{\Delta_1}{h} = \frac{d\mathcal{F}(m_t)}{dt} = -\int\left|\nabla_v \log\frac{m_t}{\hat{m}_t}(z)\right|^2 m_t(z)dz = -\int|A_t u_t|^2 m_t,$$

where the right hand side is continuous in t. The above inequality then holds for every $t > 0$. Define the 4×4 matrix

$$K_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and denote the Hilbertian norm by $\|\cdot\| = \|\cdot\|_{L^2(m_t)}$. Introduce the four-dimensional vector

$$Y_t = \big(\|A_t u_t\|, \big\|A_t^2 u_t\big\|, \|C_t u_t\|, \|C_t A_t u_t\|\big)^\mathsf{T}. \quad (2.40)$$

Then we have for almost every $t > 0$, $\lim_{h\to 0}\Delta_1/h = -Y_t^\mathsf{T} K_1 Y_t$.

Next calculate the contributions from Δ_2. Arguing as we did for Δ_1, again we have $|\Delta_2| \leqslant M_T h$. Applying the dominated convergence theorem and compute as in the proofs of [221, Lemma 32 and Theorem 18], we obtain that for almost every $t > 0$, the limit $\lim_{h\to 0}\Delta_2/h$ exists and is upper bounded by $-Y_t^\mathsf{T} K_2 Y_t$, where

$$K_2 := \begin{pmatrix} 2a - 2M_{mx}^F b & -2b & -2a & 0 \\ 0 & 2a & -2M_{mx}^F c & -4b \\ 0 & 0 & 2b & 0 \\ 0 & 0 & 0 & 2c \end{pmatrix}.$$

Step 4: Contributions from Δ_3. Now we calculate the last term

$$\Delta_3 \coloneqq I_{a,b,c}(m_t|\hat{m}_{t+h}) - I_{a,b,c}(m_t|\hat{m}_t).$$

Note that $\nabla_v \log \hat{m}_t(z) = -v$ and, by the W_2-Lipschitz continuity of the mapping $m \mapsto D_m F(m,x)$, we have

$$\begin{aligned}\big|\nabla_x \log \hat{m}_{t+h}(z) - \nabla_x \log \hat{m}_t(z)\big| &= \big|D_m F\big(m^x_{t+h},x\big) - D_m F\big(m^x_t,x\big)\big| \\ &\leqslant M^F_{mm} W_2\big(m^x_{t+h},m^x_t\big).\end{aligned}$$

So for each $z \in \mathbb{R}^{2d}$, we know that $\nabla \log \hat{m}_t(z)$ is continuous in t, and is absolutely continuous once $t \mapsto m^x_t$ is absolutely continuous with respect to the W_2 distance in the sense of [4, Definition 1.1.1]. Let us show the latter. Integrating the speed component in the Fokker–Planck equation (2.12), we obtain

$$\partial_t m^x_t + \nabla_x \cdot \big(v^x_t m^x_t\big) = 0, \tag{2.41}$$

where

$$v^x_t(x) \coloneqq \frac{\int v m_t(x,v) dv}{\int m_t(x,v) dv} = \frac{\int \nabla_v \log \frac{m_t}{\hat{m}_t}(x,v) m_t(x,v) dv}{\int m_t(x,v) dv}$$

is the average speed at the spatial point x. The L^2 norm of the vector field in the continuity equation (2.41) satisfies

$$\begin{aligned}\|v^x_t\|_{L^2(m^x_t)} &= \left(\int \left|\frac{\int \nabla_v \log \frac{m_t}{\hat{m}_t}(x,v) m_t(x,v) dv}{\int m_t(x,v) dv}\right|^2 m^x_t(x) dx\right)^{1/2} \\ &\leqslant \left(\int \left|\nabla_v \log \frac{m_t}{\hat{m}_t}(z)\right|^2 m_t(dz)\right)^{1/2} = \|A_t u_t\| \leqslant M_T,\end{aligned}$$

where the first inequality is due to Cauchy–Schwarz. Applying [4, Proposition 8.3.1] to the flow $t \mapsto m^x_t$ and its continuity equation (2.41), and using [4, Theorem 1.1.2], we obtain

$$W_2\big(m^x_{t+h},m^x_t\big) \leqslant \int_t^{t+h} \|A_s u_s\| ds \leqslant M_T h$$

for every t, h such that t and $t+h$ belong to $[0,T]$. So the mapping $t \mapsto \nabla \log \hat{m}_t(z)$ is absolutely continuous with almost everywhere derivatives satisfying

$$\begin{aligned}\partial_t \nabla_v \log \hat{m}_t(z) &= 0, \\ |\partial_t \nabla_x \log \hat{m}_t(z)| &\leqslant M^F_{mm} \|A_t u_t\| \leqslant M_T.\end{aligned}$$

Then we obtain $|\Delta_3| \leqslant M_T h$. Moreover, by the dominated convergence theorem, we have for almost every $t > 0$,

$$\begin{aligned}\lim_{h\to 0} \frac{|\Delta_3|}{|h|} &\leqslant 2M^F_{mm} \int \big(|A_t u_t(z)|, |C_t u_t(z)|\big) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} 0 \\ \|A_t u_t\| \end{pmatrix} m_t(dz) \\ &\leqslant 2M^F_{mm}\big(b\|A_t u_t\|\|A_t u_t\| + c\|A_t u_t\|\|C_t u_t\|\big) = Y_t^\top K_3 Y_t\end{aligned}$$

by applying Cauchy–Schwarz again, where

$$K_3 \coloneqq 2M^F_{mm} \begin{pmatrix} b & 0 & c & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Step 5: Hypocoercivity. Our previous bounds on Δ_k, $k = 1, 2, 3$ establish that $t \mapsto \mathcal{E}(m_t)$ is absolutely continuous (locally Lipschitz, in fact) with its almost everywhere derivative satisfying $\frac{d}{dt}\mathcal{E}(m_t) \leqslant -Y_t^\mathsf{T} K Y_t$, where K is defined by $K_1 + K_2 - K_3$ and is equal to

$$\begin{pmatrix} 1 + 2M^F_{mm}a - 2\big(M^F_{mx} + M^F_{mm}\big)b & -2b & -2a - 2M^F_{mm}c & 0 \\ 0 & 2a & -2M^F_{mx}c & -4b \\ 0 & 0 & 2b & 0 \\ 0 & 0 & 0 & 2c \end{pmatrix}.$$

As in the end of the proof of [221, Theorem 18], we can pick constants a, b, $c > 0$, depending only on M^F_{mx} and M^F_{mm}, such that $ac > b^2$ and the matrix K is a positive-definite. Let α be the smallest eigenvalue of K. Then we have

$$\begin{aligned} \frac{d\mathcal{E}(m_t)}{dt} &\leqslant -\alpha\big(\|A_t u_t\|^2 + \|C_t u_t\|^2 + \|A_t^2 u_t\|^2 + \|C_t A_t u_t\|^2\big) \\ &\leqslant -\alpha\big(\|A_t u_t\|^2 + \|C_t u_t\|^2\big) = -\alpha I(m_t|\hat{m}_t). \end{aligned}$$

Hence for every t, s such that $t \geqslant s \geqslant 0$,

$$\mathcal{E}(m_t) \leqslant \mathcal{E}(m_s) - \alpha \int_s^t I(m_u|\hat{m}_u)du. \tag{2.42}$$

Step 6: Approximation. We now show that the inequality (2.42) holds without additional assumptions on the mean field functional F and the initial value m_0.

First, suppose still that F satisfies (2.32) but no longer suppose m_0 is such that $m_0/m_\infty \in \mathcal{A}_+$. The initial value m_0 belongs to $\mathcal{P}_2(\mathbb{R}^d)$ and both $H(m_0)$ and $I(m_0)$ are finite, so thanks to Proposition 2.14, we can pick a sequence of measures $\big(m'_{n,0}\big)_{n\in\mathbb{N}}$, each of which belongs to $\mathcal{A}_+$, such that

$$\lim_{n\to\infty} W_2\big(m'_{n,0}, m_0\big) + \big|H\big(m'_{n,0}\big) - H(m_0)\big| + \big|I\big(m'_{n,0}\big) - I(m_0)\big| = 0.$$

As proved above, the inequality (2.42) holds for the flow $\big(m'_{n,t}\big)_{t\geqslant 0}$, that is,

$$\mathcal{E}\big(m'_{n,t}\big) \leqslant \mathcal{E}\big(m'_{n,0}\big) - \alpha \int_0^t I\big(m'_{n,s}|\hat{m}'_{n,s}\big)ds.$$

By the continuity with respect to the initial value of the SDE system (2.10), we have also $m'_{n,t} \to m_t$ in the weak topology of $\mathcal{P}_2$. We recall in Lemma B.1 that both the entropy and the Fisher information are lower semicontinuous with respect to the weak topology of $\mathcal{P}_2$. Taking the lower limit on both sides of the inequality above, we obtain (2.42) with $s = 0$ for the original flow $(m_t)_{t\geqslant 0}$.

Second, we no longer require F to satisfy (2.32) and set $F_k(m) = F(m \star \rho_k)$ for a sequence of smooth and symmetric mollifiers $(\rho_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$ with $\operatorname{supp}\rho_k \subset B(0, 1/k)$. The linear derivative of the regularized mean field functional reads $\frac{\delta F_k}{\delta m}(m, \cdot) = \frac{\delta F}{\delta m}(m \star \rho_k, \cdot) \star \rho_k$, and its intrinsic derivative reads $D_m F_k(m, \cdot) = D_m F(m \star \rho_k, \cdot) \star \rho_k$. Consequently,

$$\begin{aligned} |D_m F_k(m', x') - D_m F(m, x)| \leqslant M^F_{mm} W_2(m', m) + M^F_{mx}|x' - x| \\ + \frac{M^F_{mx} + M^F_{mm}}{k}. \end{aligned} \tag{2.43}$$

Moreover, $\nabla D_m F_k(m,\cdot) = \nabla D_m F(m\star\rho_k,\cdot)\star\rho_k$ and

$$\nabla^k D_m F_k(m,\cdot) = D_m F(m\star\rho_k,\cdot)\star\nabla^k\rho_k = \nabla D_m F(m\star\rho_k,\cdot)\star\nabla^{k-1}\rho_k$$

is continuous for $k\geqslant 0$ and bounded for $k\geqslant 1$. In particular F_k satisfies (2.32). Define $\mathcal{E}_k(m) = F_k(m) + \frac{1}{2}\int|v|^2 m + H(m) + I_{a,b,c}(m|\hat m)$ and here $\hat m$ should be understood as the Gibbs-type measure defined with F_k instead of F. Let $(m''_{k,t})_{t\geqslant 0}$ be the flow of measures driven by F_k with the initial value $m''_{k,0} = m_0$. Our previous result yields for every $t\geqslant 0$,

$$\mathcal{E}_k\big(m''_{k,t}\big) \leqslant \mathcal{E}_k(m_0) - \alpha\int_0^t I\big(m''_{k,s}\big|\hat m''_{k,s}\big)ds,$$

where $\hat m''_{k,s}$ is the probability measure proportional to

$$\exp\biggl(-\frac{\delta F_k}{\delta m}\bigl(m''_{k,s},x\bigr) - \frac{1}{2}|v|^2\biggr)dxdv.$$

From the bound (2.43) we deduce that $m''_{k,t}\to m_t$ in $\mathcal{P}_2$ for every $t\geqslant 0$ by the synchronous coupling result in Lemma 2.22. So taking the lower limit on both sides of the previous inequality, we obtain the inequality (2.42) with $s=0$ holds for general initial values and general mean field functionals. In particular, for every $t\geqslant 0$, the measure m_t has finite entropy and finite Fisher information. Then we apply the same argument to the flow with the initial value m_s and obtain the inequality (2.42) for general $s\geqslant 0$.

Step 7: Conclusion. Define the matrix

$$S = \begin{pmatrix} a & b \\ b & c\end{pmatrix}$$

and denote by $|S|$ its largest eigenvalue. The Fisher information satisfies for every $t\geqslant 0$,

$$\begin{aligned}
I(m_t|\hat m_t) &= \frac{1}{2}I(m_t|\hat m_t) + \frac{1}{2}I(m_t|\hat m_t)\\
&\geqslant 2\rho H(m_t|\hat m_t) + \frac{1}{2}I(m_t|\hat m_t)\\
&\geqslant 2\rho\big(\mathcal{F}(m_t) - \mathcal{F}(m_\infty)\big) + \frac{1}{2|S|}I_{a,b,c}(m_t|\hat m_t)\\
&\geqslant \biggl(2\rho\wedge\frac{1}{2|S|}\biggr)\big(\mathcal{E}(m_t) - \mathcal{E}(m_\infty)\big),
\end{aligned}$$

where on the second line we applied the uniform LSI (2.7), with ρ defined by (2.6), and on the third line we used Lemma 2.9, $m_\infty = \hat m_\infty$ and $S\preceq\lambda_2$. Applying Grönwall's lemma[2] to (2.42), we obtain the desired contractivity (2.20) with $\kappa \coloneqq \alpha\big(2\rho\wedge(2|S|)^{-1}\big)$. □

[2]The mapping $t\mapsto\mathcal{E}(m_t)$ is lower semicontinuous by Lemma B.1 and non-increasing by the inequality (2.42). So it is càdlàg. It then suffices to convolute the mapping $t\mapsto\mathcal{E}(m_t)$ by a sequence of mollifiers compactly supported in $(0,1)$, apply the classical Grönwall's lemma and take the limit.

Remark 2.17. Our Theorem 2.2 can be compared to [221, Theorem 56], where kinetic mean field Langevin dynamics with two-body interaction are studied and $O(t^{-\infty})$ entropic convergence to equilibrium is shown, under the assumption that the mean field dependence is small. This restriction is lifted by our method which leverages the functional convexity.

Remark 2.18. The regularized energy functional F_k is such that $x \mapsto D_m F_k(m,x)$ has bounded derivatives of every order. However $m \mapsto D_m F_k(m,x)$ remains only Lipschitz continuous and we are not aware of any approximation argument that allows us to obtain differentiability in the measure argument. Consequently we use still the result from [4] to treat this low regularity.

2.4.3 Particle system

In this section we study the system of particles described by the *linear* Fokker–Planck equation (2.13) and the SDE (2.11). Note that since the dynamics is linear, its wellposedness is classical and we omit its proof.

We first show that for our model we can construct hypocoercive functionals whose constants are independent of the number of particles.

Lemma 2.19 (Uniform-in-N hypocoercivity)**.** *Assume F satisfies (2.2) and there exists a measure m_∞^N satisfying (2.19) and having finite exponential moments. Let $t \mapsto m_t^N$ be a solution to the N-particle Fokker–Planck equation (2.13) in $C\big([0,T];\mathcal{P}_2(\mathbb{R}^{2dN})\big)$ whose initial value m_0^N has finite entropy and finite Fisher information. Then there exist constants a, b, c, $\alpha > 0$ depending only on M_{mx}^F and M_{mm}^F such that $ac > b^2$ and the functional*

$$\begin{aligned}\mathcal{E}^N(m^N) &\coloneqq \mathcal{F}^N(m^N) + I_{a,b,c}(m^N|m_\infty^N)\\ &\coloneqq \mathcal{F}^N(m^N) + \sum_{i=1}^N \bigg(a\int \big|\nabla_{v^i}\log h^N(\boldsymbol{z})\big|^2 m^N(d\boldsymbol{z})\\ &\qquad\qquad + 2b\int \nabla_{v^i}\log h^N(\boldsymbol{z})\cdot\nabla_{x^i}\log h^N(\boldsymbol{z}) m^N(d\boldsymbol{z})\\ &\qquad\qquad + c\int \big|\nabla_{x^i}\log h^N(\boldsymbol{z})\big|^2 m^N(d\boldsymbol{z})\bigg), \end{aligned} \tag{2.44}$$

where $h^N \coloneqq m^N/m_\infty^N$, is finite on m_t^N for $t>0$; moreover, the mapping $t\mapsto \mathcal{E}^N\big(m_t^N\big)$ satisfies

$$\mathcal{E}^N\big(m_t^N\big) \leqslant \mathcal{E}^N\big(m_s^N\big) - \alpha\int_s^t I\big(m_u^N\big|m_\infty^N\big)du \tag{2.45}$$

for every t, s such that $t \geqslant s \geqslant 0$.

Remark 2.20. The constants a, b, c are possibly different from those appearing in the proof of Theorem 2.2.

Proof. We first show that the condition (2.2) implies a bound on the second-order derivatives of $\boldsymbol{x}\mapsto U^N(\boldsymbol{x}) \coloneqq NF(\mu_{\boldsymbol{x}})$. The first-order derivatives satisfy

$$\begin{aligned}\big|\nabla_i U^N(\boldsymbol{x}) - \nabla_i U^N(\boldsymbol{x}')\big| &= \big|D_m F(\mu_{\boldsymbol{x}}, x^i) - D_m F(\mu_{\boldsymbol{x}'}, x'^i)\big|\\ &\leqslant M_{mm}^F W_2(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}'}) + M_{mx}^F|x^i - x'^i|.\end{aligned}$$

Summing over i, we obtain for every $\varepsilon > 0$,

$$\begin{aligned}\big|\nabla U^N(\boldsymbol{x})-\nabla U^N(\boldsymbol{x}')\big|^2 &\leqslant (1+\varepsilon)\big(M_{mm}^F\big)^2 N W_2^2(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}'})+(1+\varepsilon^{-1})\big(M_{mx}^F\big)^2|\boldsymbol{x}-\boldsymbol{x}'|^2\\ &\leqslant \Big((1+\varepsilon)\big(M_{mm}^F\big)^2+(1+\varepsilon^{-1})\big(M_{mx}^F\big)^2\Big)|\boldsymbol{x}-\boldsymbol{x}'|^2.\end{aligned}$$

Optimizing ε yields $\big|\nabla U^N(\boldsymbol{x})-\nabla U^N(\boldsymbol{x}')\big| \leqslant \big(M_{mm}^F+M_{mx}^F\big)|\boldsymbol{x}-\boldsymbol{x}'|$. Define

$$\big\|\nabla^2 U^N\big\|_\infty = \big\|\nabla^2 U^N\big\|_{\mathrm{op},\infty} = \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{R}^{dN}} \sup_{\boldsymbol{x}'\in\mathbb{R}^{dN}:|\boldsymbol{x}'|_2=1} \big|\nabla^2 U^N(\boldsymbol{x})\boldsymbol{x}'\big|_2.$$

From the Lipschitz bound we obtain

$$\big\|\nabla^2 U^N\big\|_{\mathrm{op},\infty} \leqslant M_{mx}^F + M_{mm}^F. \tag{2.46}$$

Now suppose there exist a constant M such that U^N satisfies

$$\boldsymbol{x}\mapsto U^N(\boldsymbol{x}) \text{ is } C^4 \qquad \text{and} \qquad \sum_{k=3}^4 \big\|\nabla^k U^N\big\|_\infty \leqslant M, \tag{2.47}$$

and that $h_0^N = m_0^N/m_\infty^N$ satisfies

$$h_0^N(\boldsymbol{z}) + \frac{1}{h_0^N(\boldsymbol{z})} + \sum_{k=1}^4 \big|\nabla^k h_t^N(\boldsymbol{z})\big| \leqslant M \tag{2.48}$$

for every $\boldsymbol{z}\in\mathbb{R}^{2dN}$. We apply Proposition 2.15 to show that under our assumptions, there exists a constant M_T such that

$$h_t^N(\boldsymbol{z}) + \frac{1}{h_t^N(\boldsymbol{z})} + \sum_{k=1}^4 \big|\nabla^k h_t^N(\boldsymbol{z})\big| \leqslant \exp\big(M_T(1+|\boldsymbol{z}|)\big) \tag{2.49}$$

for every $(t,\boldsymbol{z})\in[0,T]\times\mathbb{R}^{2dN}$ (in fact, $\boldsymbol{z}\mapsto h_t^N(\boldsymbol{z})$ remains lower and upper bounded and its up-to-fourth-order derivatives grow at most polynomially).

We denote $u_t^N = \log h_t^N$. In view of the regularity bound (2.49), we have

$$\begin{aligned}
-\frac{d\mathcal{F}^N(m_t^N)}{dt} &= \sum_{i=1}^N \int |\nabla_{v^i} u_t^N|^2 m_t^N,\\
-\frac{d}{dt}\int \big|\nabla_{v^i} u_t^N\big|^2 m_t^N &= 2\int \Big(\nabla_{x^i} u_t^N\cdot\nabla_{v^i} u_t^N + \big|\nabla_{v^i}^2 u_t^N\big|^2\\
&\qquad + \nabla_{v^i} u_t^N\cdot\nabla_{v^i} u_t^N\Big) m_t^N,\\
-\frac{d}{dt}\int \nabla_{v^i} u_t^N\cdot\nabla_{x^i} u_t^N m_t^N &= \int\Big(-\sum_{j=1}^N \nabla_{v^i} u_t^N \nabla_{ij}^2 U^N \nabla_{v^j} u_t^N + \big|\nabla_{x^i} u_t^N\big|^2\\
&\qquad + 2\nabla_{v^i}^2 u_t^N\cdot\nabla_{v^i}\nabla_{x^i} u_t^N + \nabla_{v^i} u_t^N\cdot\nabla_{x^i} u_t^N\Big) m_t^N,\\
-\frac{d}{dt}\int \nabla_{x^i} u_t^N\cdot\nabla_{x^i} u_t^N m_t^N &= \int\Big(-2\sum_{j=1}^N \nabla_{x^i} u_t^N \nabla_{ij}^2 U^N \nabla_{v^j}\nabla_{v^j} u_t^N\\
&\qquad + 2\big|\nabla_{x^i}\nabla_{v^i} u_t^N\big|^2\Big) m_t^N,
\end{aligned}$$

as is computed in [221]. Denote the Hilbertian norm by $\|\cdot\| = \|\cdot\|_{L^2(m_t^N)}$ and define the four-dimensional vector

$$Y_t^N = \big(\|\nabla_{\boldsymbol{v}} u_t^N\|, \|\nabla_{\boldsymbol{v}}^2 u_t^N\|, \|\nabla_{\boldsymbol{x}} u_t^N\|, \|\nabla_{\boldsymbol{x}}\nabla_{\boldsymbol{v}} u_t^N\|\big)^\top. \tag{2.50}$$

By Cauchy–Schwarz we have $-\frac{d}{dt}\mathcal{E}^N\big(m_t^N\big) \geqslant (Y_t^N)^\top K Y_t^N$ where

$$K \coloneqq \begin{pmatrix} 1+2a-2\big\|\nabla^2 U^N\big\|_{\mathrm{op},\infty} b & -2b & -2a & 0 \\ & 2a & -2\big\|\nabla^2 U^N\big\|_{\mathrm{op},\infty} c & -4b \\ & & 2b & 0 \\ & & & 2c \end{pmatrix},$$

where $\|\nabla^2 U^N\|_{\mathrm{op},\infty}$ is bounded by (2.46). We then apply the same argument as in the proof of Theorem 2.2 to pick a, b, c such that $ac > b^2$ and K is positive-definite with its smallest eigenvalue $\alpha > 0$. Then,

$$-\frac{d\mathcal{E}^N\big(m_t^N\big)}{dt} \geqslant (Y_t^N)^\top K Y_t^N \geqslant \alpha I\big(m_t^N \big| m_\infty^N\big),$$

from which the desired inequality (2.45) follows.

We then show the inequality (2.45) holds for general mean field functional F and initial value m_0^N. First, suppose still that U^N satisfies additionally the bound (2.47) but no longer suppose m_0^N satisfies additionally (2.48). As m_0^N has finite second moment, finite entropy and finite Fisher information, we can find a sequence of measures $(m'^N_{n,0})_{n\in\mathbb{N}}$, each of which satisfies the bound (2.48), such that

$$\lim_{n\to+\infty} W_2\big(m'^N_{n,0}, m_0^N\big) + \big|H\big(m'^N_{n,0}\big) - H\big(m'^N_0\big)\big| + \big|I\big(m'^N_{n,0}\big) - I\big(m'^N_0\big)\big| = 0,$$

by the procedure in the proof of Proposition 2.14. We have the convergence $m'^N_{n,t} \to m_t^N$ in $\mathcal{P}_2$. So taking the lower limit on both sides of

$$\mathcal{E}^N\big(m'^N_{n,t}\big) - \mathcal{E}^N\big(m'^N_{n,0}\big) + \alpha\int_0^t I\big(m'^N_{n,s}\big|m_\infty^N\big)ds \leqslant 0$$

yields (2.45) for $s = 0$, thanks to the continuity of F and the lower-semicontinuity of entropy and Fisher information with respect to the topology of $\mathcal{P}_2$, proved in Lemma B.1.

Second, we no longer suppose U^N satisfies the bound (2.47) and set

$$U_k^N = U^N \star \rho_k$$

for a sequence of smooth mollifiers $(\rho_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^{dN}$. Then U_k^N is C^4 and satisfies its second and fourth-order derivatives $\nabla^\nu U_k^N = \nabla^2 U^N \star \nabla^{\nu-2}\rho_k$ with $\nu = 3$, 4 are bounded as $\big\|\nabla^2 U^N\big\|_\infty \leqslant M_{mx}^F + M_{mm}^F$. Moreover, from the bound (2.46) on $\nabla^2 U^N$ we deduce

$$\big\|\nabla(U_k^N - U^N)\big\|_\infty \to 0 \tag{2.51}$$

and $\big\|\nabla^2 U_k^N\big\|_\infty \leqslant \big\|\nabla^2 U^N\big\|_\infty \leqslant M_{mx}^F + M_{mm}^F$. Let $\big(m''^N_{k,t}\big)_{t\geqslant 0}$ be the flow of measures driven by the regularized potential U_k^N with the initial value $m''^N_{k,0} = m_0^N$ and denote its invariant measure by $m''^N_{k,\infty}$. That is to say, $m''^N_{k,\infty}$ is the probability measure proportional to $\exp\big(-U_k^N(\boldsymbol{x}) - \frac{1}{2}|\boldsymbol{v}|^2\big)d\boldsymbol{x}d\boldsymbol{v}$. Thanks to the bound (2.51), we can apply the synchronous coupling result in Lemma 2.22 and obtain $m''^{N}_{k,t} \to m^N_t$ in $\mathcal{P}_2$ for every $t \geqslant 0$. The result obtained in the previous paragraph writes

$$H\big(m''^{N}_{k,t}\big|m''^{N}_{k,\infty}\big) + I_{a,b,c}\big(m''^{N}_{k,t}\big|m''^{N}_{k,\infty}\big) - H\big(m^N_0\big|m''^{N}_{k,\infty}\big) - I_{a,b,c}\big(m^N_0\big|m''^{N}_{k,\infty}\big) \\ + \alpha\int_0^t I\big(m''^{N}_{k,s}\big|m''^{N}_{k,\infty}\big)ds \leqslant 0$$

for every $t \geqslant 0$ and we take the lower limit on both sides to obtain (2.45) with $s = 0$ for general initial values and general mean field functional. In particular, this implies for every $t \geqslant 0$, m^N_t has finite entropy and finite Fisher information. Then we apply the same argument to the flow with m^N_s as the initial value and obtain (2.45) for general $s \geqslant 0$. □

Remark 2.21. If we additionally assume a uniform-in-N LSI for m^N_∞, then we can directly establish

$$\frac{d\mathcal{E}_N\big(m^N_t\big)}{dt} \leqslant -\kappa\mathcal{E}_N\big(m^N_t\big),$$

for a constant $\kappa > 0$ independent of N. This approach has been explored in a number of previous works. We do not impose such an assumption or sufficient conditions for it, as they often requires the mean field interaction to be small enough or (semi-)convex enough, excluding the application to neural networks in Section 2.3.

We then give the proof of Theorem 2.3. The method of proof is similar to Theorem 1.12 and we only need to take into account of the additional kinetic terms. We give a complete proof only for the sake of self-containedness.

Proof of Theorem 2.3. We pick the positive constants a, b, c, α depending only on M^F_{mx} and M^F_{mm} such that $ac > b^2$ and (2.45) holds for every $t \geqslant 0$, according to Lemma 2.19. Then, as in the proof of Theorem 1.12, we will establish a lower bound of the relative Fisher information $I_t \coloneqq I\big(m^N_t\big|m^N_\infty\big)$ in order to obtain the desired result.

Step 1: Regularity of conditional distribution. By local hypoelliptic positivity (see e.g. [221, Theorem A.19 and Corollary A.21]), we know that for every $t > 0$ and every $\boldsymbol{z} \in \mathbb{R}^{2dN}$, $m^N_t(\boldsymbol{z}) > 0$. Let $i \in \{1, \dots, N\}$. Define the marginal density $m^{N,-i}_t(\boldsymbol{z}^{-i}) = \int m^N_t(\boldsymbol{z})dz^i$, which is strictly positive by the local positivity of m^N_t and is lower semicontinuous by Fatou's lemma. By the Fubini theorem, we have $\int m^{N,-i}_t(\boldsymbol{z}^{-i})d\boldsymbol{z}^{-i} = 1$. Together with the lower semicontinuity, we obtain that $m^{N,-i}_t(\boldsymbol{z}^{-i})$ is finite everywhere. We are therefore able to define the conditional probability density

$$m^{N,i|-i}_t(z^i|\boldsymbol{z}^{-i}) = \frac{m^N_t(\boldsymbol{z})}{m^{N,-i}_t(\boldsymbol{z}^{-i})} = \frac{m^N_t(\boldsymbol{z})}{\int m^N_t(\boldsymbol{z})dz^i},$$

which is weakly differentiable in z^i and strictly positive everywhere. We can also define the conditional density $m^{N,i|-i}_\infty$ for the invariant measure m^N_∞, and the regularity follows directly from its explicit expression.

Step 2: Decomposing Fisher componentwise. Using the conditional distributions, we can decompose the relative Fisher information as

$$\begin{aligned}
I_t &= \int \bigg|\nabla \log \frac{m_t^N(\boldsymbol{z})}{m_\infty^N(\boldsymbol{z})}\bigg|^2 m_t^N(d\boldsymbol{z}) = \mathbb{E}\Bigg[\bigg|\nabla \log \frac{m_t^N(\boldsymbol{Z}_t)}{m_\infty^N(\boldsymbol{Z}_t)}\bigg|^2\Bigg] \\
&= \sum_{i=1}^N \mathbb{E}\Bigg[\bigg|\nabla_{z^i} \log \frac{m_t^{N,i|-i}(Z_t^i|\boldsymbol{Z}_t^{-i}) m_t^{N,-i}(\boldsymbol{Z}_t^{-i})}{m_\infty^N(\boldsymbol{Z}_t)}\bigg|^2\Bigg] \\
&= \sum_{i=1}^N \mathbb{E}\Bigg[\bigg|\nabla_{z^i} \log \frac{m_t^{N,i|-i}(Z_t^i|\boldsymbol{Z}_t^{-i})}{m_\infty^N(\boldsymbol{Z}_t)}\bigg|^2\Bigg] \\
&= \sum_{i=1}^N \mathbb{E}\Big[\Big|\nabla_{z^i} \log m_t^{N,i|-i}(Z_t^i|\boldsymbol{Z}_t^{-i}) + D_m F(\mu_{\boldsymbol{X}_t}, X_t^i) + V_t^i\Big|^2\Big].
\end{aligned}$$

Step 3: Change of empirical measure and componentwise LSI. We replace the empirical measure $\mu_{\boldsymbol{x}}$ in $D_m F$ by $\mu_{\boldsymbol{x}^{-i}}$. Define the difference $\delta_1^i(\boldsymbol{x}; y) = D_m F(\mu_{\boldsymbol{x}}, y) - D_m F(\mu_{\boldsymbol{x}^{-i}}, y)$ and denote by $\hat{\mu}_{\boldsymbol{x}^{-i}}$ the probability on $\mathbb{R}^{2d}$ such that

$$\hat{\mu}_{\boldsymbol{x}^{-i}}(dxdv) \propto \exp\bigg(-\frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}^{-i}}, x) - \frac{1}{2}|v|^2\bigg) dxdv.$$

For every $\varepsilon \in (0,1)$, the Fisher information satisfies

$$\begin{aligned}
I_t &= \sum_{i=1}^N \mathbb{E}\Big[\Big|\nabla_{x^i} \log m_t^{N,i|-i}(Z_t^i|\boldsymbol{Z}_t^{-i}) + D_m F(\mu_{\boldsymbol{X}_t^{-i}}, X_t^i) + V_t^i + \delta_1^i(\boldsymbol{X}_t; X_t^i)\Big|^2\Big] \\
&\geqslant \sum_{i=1}^N \mathbb{E}\Bigg[(1-\varepsilon)\Big|\nabla_{x^i} \log m_t^{N,i|-i}(X_t^i|\boldsymbol{X}_t^{-i}) + D_m F(\mu_{\boldsymbol{X}_t^{-i}}, X_t^i) + V_t^i\Big|^2 \\
&\qquad\qquad - (\varepsilon^{-1}-1)\big|\delta_1^i(\boldsymbol{X}_t; X_t^i)\big|^2\Bigg] \\
&= (1-\varepsilon) \sum_{i=1}^N \mathbb{E}\Big[I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big|\hat{\mu}_{\boldsymbol{X}_t^{-i}}\Big)\Big] - (\varepsilon^{-1}-1) \sum_{i=1}^N \mathbb{E}\Big[\big|\delta_1^i(\boldsymbol{X}_t; X_t^i)\big|^2\Big],
\end{aligned}$$

where we used the elementary inequality $(a+b)^2 \geqslant (1-\varepsilon)|a|^2 - (\varepsilon^{-1}-1)|b|^2$. Define the first error

$$\Delta_1 \coloneqq \sum_{i=1}^N \mathbb{E}\Big[\big|\delta_1^i(\boldsymbol{X}_t; X_t^i)\big|^2\Big] \coloneqq \sum_{i=1}^N \mathbb{E}\Big[\big|D_m F(\mu_{\boldsymbol{X}_t}, X_t^i) - D_m F(\mu_{\boldsymbol{X}_t^{-i}}, X_t^i)\big|^2\Big]. \tag{2.52}$$

The previous inequality writes

$$I_t \geqslant (1-\varepsilon) \sum_{i=1}^N \mathbb{E}\Big[I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big|\hat{\mu}_{\boldsymbol{X}_t^{-i}}\Big)\Big] - (\varepsilon^{-1}-1)\Delta_1. \tag{2.53}$$

We apply the uniform ρ-log-Sobolev inequality (2.7) for $\hat{\mu}_{\boldsymbol{X}_t^i}$ with ρ defined by (2.6) and obtain

$$\begin{aligned}
\frac{1}{4\rho} I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big|\hat{\mu}_{\boldsymbol{X}_t^{-i}}\Big) &\geqslant H\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big|\hat{\mu}_{\boldsymbol{X}_t^{-i}}\Big) \\
= \int \bigg(\log m_t^{N,i|-i}(x^i|\boldsymbol{Z}_t^{-i}) &+ \frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t^{-i}}, x^i) + \frac{1}{2}|v^i|^2\bigg) m_t^{N,i|-i}(dz^i|\boldsymbol{Z}_t^{-i}) \\
&\qquad + \log Z(\hat{\mu}_{\boldsymbol{X}_t^{-i}}),
\end{aligned}$$

where the last quantity is the normalization factor

$$Z(\hat{\mu}_{\boldsymbol{X}_t^{-i}}) \coloneqq \int \exp\Big(-\frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t^{-i}}, x) - \frac{1}{2}|v|^2\Big) dx dv.$$

Then we apply Jensen's inequality to $\log Z(\hat{\mu}_{\boldsymbol{x}^{-i}})$ to obtain

$$\log Z(\hat{\mu}_{\boldsymbol{X}_t^{-i}}) \geqslant -\int \Big(\frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t^{-i}}, x^i) + \frac{1}{2}|v^i|^2\Big) m_\infty(dz^i) - \int m_\infty(z^i)\log m_\infty(z^i) dz^i.$$

Chaining the previous two inequalities and summing over i, we have

$$\begin{aligned}\frac{1}{4\rho}\sum_{i=1}^N I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big|\hat{\mu}_{\boldsymbol{X}_t^{-i}}\Big) \geqslant \sum_{i=1}^N \Big[\int\Big(\frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t^{-i}}, x^i) + \frac{1}{2}|v^i|^2\Big)\\ \Big(m_t^{N,i|-i}(dz^i|\boldsymbol{Z}_t^{-i}) - m_\infty(dz^i)\Big) + H\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big) - H(m_\infty)\Big]. \quad (2.54)\end{aligned}$$

Step 4: Another change of empirical measure. We are going to replace $\mu_{\boldsymbol{x}^{-i}}$ by $\mu_{\boldsymbol{x}}$ in (2.54). Define $\delta_2^i(\boldsymbol{x}; y) \coloneqq \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}^{-i}}, y) - \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}}, y)$ and the second error

$$\Delta_2 \coloneqq \sum_{i=1}^N \int \delta_2^i(\boldsymbol{x}; x^i) m_t^N(d\boldsymbol{z}) - \sum_{i=1}^N \iint \delta_2^i(\boldsymbol{x}; x') m_\infty(dz') m_t^N(d\boldsymbol{z}). \quad (2.55)$$

Taking expectations on both sides of (2.54), we obtain

$$\begin{aligned}&\frac{1}{4\rho}\sum_{i=1}^N \mathbb{E}\Big[I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big|\hat{\mu}_{\boldsymbol{X}_t^{-i}}\Big)\Big]\\ &\geqslant N\,\mathbb{E}\Big[\int\Big(\frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t}, x) + \frac{1}{2}|v|^2\Big)(\mu_{\boldsymbol{Z}_t} - m_\infty)(dz)\Big] + \sum_{i=1}^N \mathbb{E}\Big[H\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big)\Big]\\ &\qquad - NH(m_\infty) + \Delta_2. \quad (2.56)\end{aligned}$$

Thanks to the convexity of F, the first term satisfies the tangent inequality

$$\begin{aligned}&N\,\mathbb{E}\Big[\int\Big(\frac{\delta F}{\delta m}(\mu_{\boldsymbol{X}_t}, x) + \frac{1}{2}|v|^2\Big)(\mu_{\boldsymbol{Z}_t} - m_\infty)(dz)\Big]\\ &\quad\geqslant N\,\mathbb{E}\big[F(\mu_{\boldsymbol{X}_t}) - F(m_\infty^x)\big] + \frac{1}{2}\int |\boldsymbol{v}|^2 m_t^N(d\boldsymbol{z}) - \frac{N}{2}\int |v^2| m_\infty(dz)\\ &\quad= F^N(m_t^N) - NF(m_\infty) + \frac{1}{2}\int |\boldsymbol{v}|^2 m_t^N(d\boldsymbol{z}) - \frac{N}{2}\int |v^2| m_\infty(dz). \quad (2.57)\end{aligned}$$

For the second term we apply the information inequality (2.30) to obtain

$$\sum_{i=1}^N \mathbb{E}^{-i}\Big[H\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big)\Big] \geqslant H(m_t^N).$$

Hence,

$$\sum_{i=1}^N \mathbb{E}\Big[I\Big(m_t^{N,i|-i}(\cdot|\boldsymbol{Z}_t^{-i})\Big|\hat{\mu}_{\boldsymbol{X}_t^{-i}}\Big)\Big] \geqslant 4\rho\Big(\mathcal{F}^N(m_t^N) - N\mathcal{F}(m_\infty) + \Delta_2\Big)$$

by the definition of free energies $\mathcal{F}(m) = F(m) + H(m) + \frac{1}{2}\int |v|^2 m$, $\mathcal{F}^N(m^N) = F^N(m^N) + H(m^N) + \frac{1}{2}\int |\boldsymbol{v}|^2 m^N$. Using (2.53), we obtain

$$I_t = I\big(m_t^N \big| m_\infty^N\big) \geqslant 4\rho(1-\varepsilon)\Big(\mathcal{F}^N\big(m_t^N\big) - N\mathcal{F}(m_\infty) + \Delta_2\Big) - (\varepsilon^{-1}-1)\Delta_1. \quad (2.58)$$

Step 5: Bounding the errors Δ_1, Δ_2. The transport plan between $\mu_{\boldsymbol{x}}$ and $\mu_{\boldsymbol{x}^{-i}}$

$$\pi^i = \frac{1}{N}\sum_{j\neq i}\delta_{(x^j,x^j)} + \frac{1}{N(N-1)}\sum_{j\neq i}\delta_{(x^j,x^i)} \quad (2.59)$$

gives the bound $W_1(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}}) \leqslant \frac{1}{N(N-1)}\sum_{j\neq i}|x^j - x^i|$. We use this transport plan to bound the errors Δ_1, Δ_2.

Let us treat the first error Δ_1. Since $m \mapsto D_m F(m,x)$ is M_{mm}^F-Lipschitz continuous in W_2 metric, we have

$$\big|\delta_1^i(\boldsymbol{x};y)\big| \leqslant M_{mm}^F W_2(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}}).$$

Under the L^2-optimal transport plan $\operatorname{Law}\big((Z_t^i)_{i=1}^N, (\tilde{Z}_\infty^i)_{i=1}^N\big) \in \Pi(m_t^N, m_\infty^{\otimes N})$ we have

$$\begin{aligned}
\Delta_1 &= \sum_{i=1}^N \mathbb{E}\Big[\big|\delta_1^i(\boldsymbol{X}_t;X_t^i)\big|^2\Big] \leqslant \big(M_{mm}^F\big)^2 \sum_{i=1}^N \mathbb{E}\big[W_1^2\big(\mu_{\boldsymbol{X}_t},\mu_{\boldsymbol{X}_t^{-i}}\big)\big] \\
&\leqslant \frac{\big(M_{mm}^F\big)^2}{N(N-1)}\,\mathbb{E}\Bigg[\sum_{\substack{1\leqslant i,j\leqslant N\\ i\neq j}} \big|X_t^j - X_t^i\big|^2\Bigg] \\
&\leqslant \frac{3\big(M_{mm}^F\big)^2}{N(N-1)}\,\mathbb{E}\Bigg[\sum_{\substack{1\leqslant i,j\leqslant N\\ i\neq j}} \Big(\big|X_t^i - \tilde{X}_\infty^i\big|^2 + \big|\tilde{X}_\infty^i - \tilde{X}_\infty^j\big|^2 + \big|X_t^j - \tilde{X}_\infty^j\big|^2\Big)\Bigg] \\
&\leqslant \frac{3\big(M_{mm}^F\big)^2}{N(N-1)}\Bigg(2(N-1)\,\mathbb{E}\Bigg[\sum_{i=1}^N\big|X_t^i - \tilde{X}_\infty^i\big|^2\Bigg] + N(N-1)\,\mathbb{E}\Big[\big|\tilde{X}_\infty^1 - \tilde{X}_\infty^2\big|^2\Big]\Bigg).
\end{aligned}$$

The first term $\mathbb{E}\big[\sum_{i=1}^N |X_t^i - \tilde{X}_\infty^i|^2\big]$, being only the transport cost in the X directions, is bounded by the Wasserstein distance $W_2^2\big(m_t^N, m_\infty^{\otimes N}\big)$, while the second $\mathbb{E}\big[|\tilde{X}_\infty^1 - \tilde{X}_\infty^2|^2\big]$ equals $2\operatorname{Var} m_\infty^x$. Hence the first error satisfies the bound

$$\Delta_1 \leqslant 6\big(M_{mm}^F\big)^2\bigg(\frac{1}{N}W_2^2\big(m_t^N, m_\infty^{\otimes N}\big) + \operatorname{Var} m_\infty\bigg). \quad (2.60)$$

Now treat the second error Δ_2. The Lipschitz constant of $y \mapsto \delta_2^i(\boldsymbol{x};y) = \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}^{-i}},y) - \frac{\delta F}{\delta m}(\mu_{\boldsymbol{x}},y)$ is controlled by

$$\big|\nabla_y\delta_2^i(\boldsymbol{x};y)\big| = |D_mF(\mu_{\boldsymbol{x}},y) - D_mF(\mu_{\boldsymbol{x}^{-i}},y)| \leqslant M_{mm}^F W_1(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}}).$$

Hence $\big|\delta_2^i(\boldsymbol{x};y) - \delta_2^i(\boldsymbol{x};y')\big| \leqslant M_{mm}^F W_1(\mu_{\boldsymbol{x}},\mu_{\boldsymbol{x}^{-i}})|y-y'|$. Use Fubini's theorem to first integrate z' in the definition of the second error (2.55) and let $\tilde{Z}_\infty'$ be independent from $\boldsymbol{Z}_t$. We obtain

$$
\begin{aligned}
|\Delta_2| &\leqslant \sum_{i=1}^N \int \bigg(\int \big|\delta_2^i(\boldsymbol{x}; x^i) - \delta_2^i(\boldsymbol{x}; x')\big| m_\infty(dz') \bigg) m_t^N(d\boldsymbol{z}) \\
&\leqslant \sum_{i=1}^N \iint M_{mm}^F W_1(\mu_{\boldsymbol{x}}, \mu_{\boldsymbol{x}^{-i}}) |x' - x^i| m_\infty(dz') m_t^N(d\boldsymbol{z}) \\
&\leqslant \sum_{i=1}^N \iint \frac{M_{mm}^F}{N(N-1)} \sum_{j=1,\, j\neq i}^N |x^i - x^j||x' - x^i| m_\infty(dz') m_t^N(d\boldsymbol{z}) \\
&\leqslant \frac{M_{mm}^F}{2N(N-1)} \sum_{i=1}^N \iint \sum_{j=1,\, j\neq i}^N \big(|x^i - x^j|^2 + |x' - x^i|^2\big) m_\infty(dz') m_t^N(d\boldsymbol{z}) \\
&\leqslant \frac{M_{mm}^F}{2N(N-1)} \Bigg(\sum_{\substack{i,j=1\\ i\neq j}}^N \mathbb{E}\Big[\big|X_t^i - X_t^j\big|^2\Big] + (N-1) \sum_{i=1}^N \mathbb{E}\Big[\big|X_t^i - \tilde{X}'_\infty\big|^2\Big] \Bigg).
\end{aligned}
$$

Using the same method we used for Δ_1, we control the first term by

$$
\sum_{\substack{i,j=1\\ i\neq j}}^N \mathbb{E}\Big[\big|X_t^i - X_t^j\big|^2\Big] \leqslant 6N(N-1) \bigg(\frac{1}{N} W_2^2\big(m_t^N, m_\infty^{\otimes N}\big) + \operatorname{Var} m_\infty^x \bigg).
$$

For the second term we work again under the L^2-optimal plan

$$
\operatorname{Law}\big((Z_t^i)_{i=1}^N, (\tilde{Z}_\infty^i)_{i=1}^N\big) \in \Pi(m_t^N, m_\infty^{\otimes N})
$$

and let $\tilde{Z}'_\infty$ remain independent from the other variables. We have

$$
\begin{aligned}
\sum_{i=1}^N \mathbb{E}\Big[\big|X_t^i - \tilde{X}'_\infty\big|^2\Big] &\leqslant 2 \sum_{i=1}^N \Big(\mathbb{E}\Big[\big|X_t^i - \tilde{X}_\infty^i\big|^2\Big] + \mathbb{E}\Big[\big|\tilde{X}_\infty^i - \tilde{X}'_\infty\big|^2\Big] \Big) \\
&\leqslant 2N \bigg(\frac{1}{N} W_2^2\big(m_t^N, m_\infty^{\otimes N}\big) + 2 \operatorname{Var} m_\infty^x \bigg).
\end{aligned}
$$

As a result,

$$
|\Delta_2| \leqslant M_{mm}^F \bigg(\frac{4}{N} W_2^2\big(m_t^N, m_\infty^{\otimes N}\big) + 5 \operatorname{Var} m_\infty^x \bigg). \tag{2.61}
$$

Step 6: Conclusion. Inserting the bounds on the errors (2.60), (2.61) to the lower bound of Fisher information (2.58), we obtain

$$
\begin{aligned}
I\big(m_t^N \big| m_\infty^N\big) \geqslant{}& 4\rho(1-\varepsilon) \Big(\mathcal{F}^N\big(m_t^N\big) - N\mathcal{F}(m_\infty) \Big) \\
&- \Big(16\rho M_{mm}^F + 6(\varepsilon^{-1} - 1)\big(M_{mm}^F\big)^2 \Big) \frac{1}{N} W_2^2\big(m_t^N, m_\infty^{\otimes N}\big) \\
&- \Big(20\rho M_{mm}^F + 6(\varepsilon^{-1} - 1)\big(M_{mm}^F\big)^2 \Big) \operatorname{Var} m_\infty^x.
\end{aligned}
$$

Thanks to the Poincaré inequality (2.8) for $m_\infty = \hat{m}_\infty$, its spatial variance satisfies

$$
2\rho \operatorname{Var}_{m_\infty}(x^i) \leqslant \mathbb{E}_{m_\infty}\big[|\nabla x^i|^2\big] = 1. \tag{2.62}
$$

So $\operatorname{Var} m_\infty^x = \sum_{i=1}^d \operatorname{Var}_{m_\infty}(x^i) \leqslant d/2\rho$. Using the T_2-transport inequality (2.9) for $m_\infty^{\otimes N}$ and the entropy sandwich Lemma 2.10 we bound the transport cost by

$$W_2^2\big(m_t^N, m_\infty^{\otimes N}\big) \leqslant \frac{1}{\rho} H\big(m_t^N \big| m_\infty^{\otimes N}\big) \leqslant \frac{1}{\rho}\Big(\mathcal{F}^N\big(m_t^N\big) - N\mathcal{F}(m_t)\Big).$$

In the end we obtain $\mathcal{E}^N\big(m_T^N\big) \leqslant \mathcal{E}^N\big(m_s^N\big) - \alpha \int_s^T I_t dt$ where

$$\begin{aligned} I_t &= \frac{1}{2} I\big(m_t^N \big| m_\infty^N\big) + \frac{1}{2} I\big(m_t^N \big| m_\infty^N\big) \\ &\geqslant \frac{1}{2}\bigg[4(1-\varepsilon)\rho - \frac{M_{mm}^F}{N}\bigg(16 + 6(\varepsilon^{-1}-1)\frac{M_{mm}^F}{\rho}\bigg)\bigg]\Big(\mathcal{F}^N\big(m_t^N\big) - N\mathcal{F}(m_\infty)\Big) \\ &\qquad + \frac{1}{2} I\big(m_t^N \big| m_\infty^N\big) - \frac{dM_{mm}^F}{2\rho}\big(10\rho + 3(\varepsilon^{-1}-1)M_{mm}^F\big). \end{aligned}$$

We conclude by applying Grönwall's lemma, as in the end of the proof of Theorem 2.2. □

2.5 Short-time behaviors and propagation of chaos

Our proof of the main theorem on the uniform-in-time propagation of chaos (Theorem 2.6) relies on the exponential convergence in Theorems 2.2 and 2.3, where the initial conditions are required to have finite entropy and finite Fisher information. We aim to demonstrate in this section that the non-linear kinetic Langevin dynamics exhibits the same regularization effects in short time as the linear ones, where the contributions from the non-linearity can be controlled. We will first show the short-time Wasserstein propagation of chaos using synchronous coupling. Then we adapt the regularization results for the linear dynamics to our setting and show that for measure initial values of finite second moment, the entropy and the Fisher information are finite for the flow at every positive time, where the short-time Wasserstein propagation of chaos also plays a role. Finally we combine all the estimates obtained to derive Theorem 2.6.

2.5.1 Synchronous coupling

We first show a lemma where synchronous coupling is applied to general McKean–Vlasov diffusions. This lemma is also used to justify the approximation arguments in the proof of Theorems 2.2 and 2.3.

Lemma 2.22. *Let $T > 0$ and β, $\beta' : [0,T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ be measurable and uniformly Lipschitz continuous in the last two variables and σ be a $d \times d$ real matrix. Suppose the integral $\int_0^T (|\beta(t,\delta_0,0)| + |\beta'(t,\delta_0,0)|)dt$ is finite. Let $(Z_t)_{t\in[0,T]}$, $(Z'_t)_{t\in[0,T]}$ be respective solutions to*

$$\begin{aligned} dZ_t &= \beta\big(t, \operatorname{Law}(Z_t), Z_t\big)dt + \sigma dW_t, \\ dZ'_t &= \beta'\big(t, \operatorname{Law}(Z'_t), Z'_t\big)dt + \sigma dW'_t, \end{aligned}$$

where W, W' are d-dimensional Brownians. If there exist constants M_m, M_z and a progressively measurable $\delta : \Omega \times [0,T] \to \mathbb{R}$ such that for every $t \in [0,T]$, every m, $m' \in \mathcal{P}_2(\mathbb{R}^d)$ and every x, $x' \in \mathbb{R}^d$,

$$|\beta(t,m,Z_t) - \beta'(t,m',Z'_t)| \leqslant M_m W_2(m,m') + M_z|Z_t - Z'_t| + \delta_t \tag{2.63}$$

almost surely, then for every $t \in [0,T]$,

$$W_2^2\big(\mathrm{Law}(Z_t), \mathrm{Law}(Z_t')\big) \leqslant e^{(2M_m+2M_z)t+1} W_2^2\big(\mathrm{Law}(Z_0), \mathrm{Law}(Z_0')\big) \\ + e^2 t \int_0^t e^{(2M_m+2M_z)(t-s)}\, \mathbb{E}\big[\delta_s^2\big] ds.$$

Proof. From the uniformly Lipschitz continuity of b and b' we have the uniqueness in law and the existence of strong solution for both diffusions. So we can construct $(Z_t, Z_t')_{t\in[0,T]}$ such that they share the same Brownian motion and satisfy

$$\mathbb{E}\big[|Z_0 - Z_0'|^2\big] = W_2^2\big(\mathrm{Law}(Z_0), \mathrm{Law}(Z_0')\big).$$

Consequently,

$$d(Z_s - Z_s') = \big[b\big(s, \mathrm{Law}(Z_t), Z_s\big) - b'\big(s, \mathrm{Law}(Z_s'), Z_s'\big)\big] dt$$

and by Itō's formula,

$$d|Z_s - Z_s'|^2 = 2(Z_s - Z_s') \cdot \big[\beta\big(s, \mathrm{Law}(Z_s), Z_s\big) - \beta'\big(s, \mathrm{Law}(Z_s'), Z_s'\big)\big] ds.$$

By (2.63) we have

$$\big|\beta\big(s, \mathrm{Law}(Z_s), Z_s\big) - \beta'\big(s, \mathrm{Law}(Z_s), Z_s\big)\big| \leqslant M_m W_2\big(\mathrm{Law}(Z_s), \mathrm{Law}(Z_s')\big) \\ + M_z|Z_s - Z_s'| + \delta_s \leqslant M_m\, \mathbb{E}\big[|Z_s - Z_s'|^2\big]^{1/2} + M_z|Z_s - Z_s'| + \delta_s.$$

Hence

$$\frac{1}{2} d|Z_s - Z_s'|^2 \leqslant M_z|Z_s - Z_s'|^2 + M_m\, \mathbb{E}\big[|Z_s - Z_s'|^2\big]^{1/2} |Z_s - Z_s'| + |Z_s - Z_s'|\delta_s.$$

By Cauchy–Schwarz,

$$d|Z_s - Z_s'|^2 \leqslant (2M_z + M_m + t^{-1})|Z_s - Z_s'|^2 + M_m\, \mathbb{E}\big[|Z_s - Z_s'|^2\big] + t\delta_s^2.$$

Taking expectations on both sides and applying Grönwall's lemma, we obtain

$$\mathbb{E}\big[|Z_t - Z_t'|^2\big] = e^{2(M_z+M_m)t+1}\, \mathbb{E}\big[|Z_0 - Z_0'|^2\big] + t \int_0^t e^{2(M_z+M_m+t^{-1})(t-s)}\, \mathbb{E}\big[\delta_s^2\big] ds \\ \leqslant e^{2(M_z+M_m)t+1}\, \mathbb{E}\big[|Z_0 - Z_0'|^2\big] + e^2 t \int_0^t e^{2(M_z+M_m)(t-s)}\, \mathbb{E}\big[\delta_s^2\big] ds,$$

from which the desired inequality follows. □

Since the finite-time propagation of chaos does not depend on the gradient structure of the diffusions, we introduce a more general setting. Let $b : \mathcal{P}_2(\mathbb{R}^{2d}) \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a mapping that is Lipschitz in space and velocity: there exist positive constants M_x^b, M_v^b such that

$$\forall x, x', v, v' \in \mathbb{R}^d,\ \forall m \in \mathcal{P}_2(\mathbb{R}^{2d}), \\ |b(m,x,v) - b(m,x',v')| \leqslant M_x^b|x - x'| + M_v^b|v - v'|. \quad (2.64)$$

We suppose also that the functional derivatives $\frac{\delta b}{\delta m}, \frac{\delta^2 b}{\delta m^2}$ exist with the following bounds: there exist positive constants M_m^b, M_{mm}^b such that

$$\forall m \in \mathcal{P}_2(\mathbb{R}^{2d}),\ \forall z, z' \in \mathbb{R}^{2d}, \quad |D_m b(m, z, z')|_{\mathrm{op}} \leqslant M_m^b, \tag{2.65}$$

and

$$\begin{aligned}&\forall m, m' \in \mathcal{P}_2(\mathbb{R}^{2d}),\ \forall z \in \mathbb{R}^{2d},\\ &\left|\iint \left[\frac{\delta^2 b}{\delta m^2}(m', z, z', z') - \frac{\delta^2 b}{\delta m^2}(m', z, z', z'')\right] m(dz') m(dz'')\right| \leqslant M_{mm}^b.\end{aligned} \tag{2.66}$$

We consider the following mean field dynamics:

$$\begin{aligned} dX_t &= V_t dt, \\ dV_t &= b\big(\mathrm{Law}(X_t, V_t), X_t, V_t\big) dt + \sqrt{2} dW_t, \end{aligned} \tag{2.67}$$

and the corresponding particle system:

$$\begin{aligned} &dX_t^i = V_t^i dt, \\ &dV_t^i = b\big(\mu_{(\boldsymbol{X}_t, \boldsymbol{V}_t)}, X_t^i, V_t^i\big) dt + \sqrt{2} dW_t^i, \quad \text{where } \mu_{(\boldsymbol{X}_t, \boldsymbol{V}_t)} = \frac{1}{N}\sum_{i=1}^N \delta_{(X_t^i, V_t^i)}, \end{aligned} \tag{2.68}$$

and $i = 1, \ldots, N$. In both equations W_t, W_t^i are standard Brownians and $(W_t^i)_{i=1}^N$ are independent from each other. The dynamics (2.67), (2.68) are well defined globally in time thanks to the Lipschitz continuity (2.64) and we denote by P and P^N the respective associated semigroups. That is to say, if (X_t, V_t) solves (2.67) and $\mathrm{Law}(X_0, V_0) = \mu$, then $P_t^* \mu := \mathrm{Law}(X_t, V_t)$ and $(P_t f)(\mu) := \langle f, (P_t)^* \mu \rangle$ for bounded measurable $f : \mathbb{R}^{2d} \to \mathbb{R}$; if $(X_t^i, V_t^i)_{i=1}^N$ solves (2.68) and $\mathrm{Law}(\boldsymbol{X}_0, \boldsymbol{V}_0) = \mu^N$, then $\big(P_t^N\big)^* \mu^N := \mathrm{Law}(\boldsymbol{X}_t, \boldsymbol{V}_t)$ and $\big(P_t^N f^N\big)(\mu^N) := \langle f^N, \big(P_t^N\big)^* \mu^N \rangle$ for bounded measurable $f^N : \mathbb{R}^{2dN} \to \mathbb{R}$. We also define the tensor product of the mean field semigroup: $\big(P_t^{\otimes N} f^N\big)(\mu) := \langle f^N, (P_t^* \mu)^{\otimes N} \rangle$.

Using the previous Lemma 2.22 as a building block, we now show the finite-time propagation of chaos result.

Proposition 2.23 (Finite-time propagation of chaos)**.** *Assume b satisfies (2.64) and (2.65), (2.66), and let $N \geqslant 2$. Then there exist a positive constant C depending on M_x^b, M_v^b, M_m^b and M_{mm}^b such that for every $m \in \mathcal{P}_2(\mathbb{R}^{2d})$ and $m^N \in \mathcal{P}_2(\mathbb{R}^{2dN})$, and every $T \geqslant 0$,*

$$W_2^2\Big(\big(P_T^N\big)^* m^N, (P_T^* m)^{\otimes N}\Big) \leqslant C e^{CT} W_2^2\big(m^N, m^{\otimes N}\big) + C(e^{CT} - 1)(\mathrm{Var}\, m_0 + d). \tag{2.69}$$

Proof. We apply Lemma 2.22 with

$$\begin{aligned} \beta^i(t, m^N, \boldsymbol{z}) &= \big(v^i, b(\mu_{(\boldsymbol{x}, \boldsymbol{v})}, x^i, v^i)\big)^\top, \\ \beta'^i(t, m^N, \boldsymbol{z}) &= \big(v^i, b(P_t^* m, x^i, v^i)\big)^\top, \\ \delta_t^2 &:= \sum_{i=1}^N \big|\delta_t^i\big|^2 := \sum_{i=1}^N \big|b\big(\mu_{(\boldsymbol{X}_t, \boldsymbol{V}_t)}, X_t^i, V_t^i\big) - b\big(P_t^* m, X_t^i, V_t^i\big)\big|^2, \end{aligned}$$

and $M_z \coloneqq \sqrt{2}M_x^b \vee \sqrt{2(M_v^b)^2+1}$, $M_m \coloneqq 0$. We then obtain

$$W_2^2\Big(\big(P_t^N\big)^* m^N, (P_t^* m)^{\otimes N}\Big) \leqslant e^{2M_z t+1} W_2^2\big(m_0^N, m_0^{\otimes N}\big) + e^2 t \int_0^t e^{2M_z(t-s)}\, \mathbb{E}\big[\delta_t^2\big] ds.$$

So it remains to bound $\mathbb{E}[\delta_t^2]$. By enlarging the underlying probability space, we construct the random variable $\tilde{\boldsymbol{Z}}_t = \big(\tilde{\boldsymbol{X}}_t', \tilde{\boldsymbol{V}}_t'\big) \sim (P_t^* m)^{\otimes N}$ such that

$$\sum_{i=1}^N \mathbb{E}\Big[\big|Z_t^i - \tilde{Z}_t'^i\big|^2\Big] = W_2^2\Big(\big(P_t^N\big)^* m^N, (P_t^* m)^{\otimes N}\Big).$$

This implies in particular

$$\mathbb{E}\big[W_2^2\big(\mu_{(\boldsymbol{X}_t,\boldsymbol{V}_t)}, \mu_{(\tilde{\boldsymbol{X}}_t', \tilde{\boldsymbol{V}}_t')}\big)\big] \leqslant \frac{1}{N} W_2^2\big(m_t^N, m_t^{\otimes N}\big). \tag{2.70}$$

For each i, we decompose

$$\begin{aligned}\delta_t^i = \Big(b\big(P_t^* m, Z_t^i\big) - b\big(\mu_{\tilde{\boldsymbol{Z}}_t'^{-i}}, Z_t^i\big)\Big) + \Big(b\big(\mu_{\tilde{\boldsymbol{Z}}_t'^{-i}}, Z_t^i\big) - b\big(\mu_{\tilde{\boldsymbol{Z}}_t'}, Z_t^i\big)\Big) \\ + \Big(b\big(\mu_{\tilde{\boldsymbol{Z}}_t'}, Z_t^i\big) - b\big(\mu_{\boldsymbol{Z}_t}, Z_t^i\big)\Big) \eqqcolon \text{(I)} + \text{(II)} + \text{(III)}.\end{aligned}$$

According to the assumption (2.4) we can apply Lemma B.3 to the first term and obtain

$$\mathbb{E}\big[\text{(I)}^2\big] = \mathbb{E}\Big[\mathbb{E}\big[\text{(I)}^2\big|Z_t^i\big]\Big] \leqslant \frac{\big(M_m^b\big)^2 \operatorname{Var} m_t}{N-1} + \frac{M_{mm}^b}{4(N-1)^2}.$$

We then bound the second term by the M_m^b-Lipschitz continuity:

$$\begin{aligned}\mathbb{E}\big[\text{(II)}^2\big] &\leqslant \big(M_m^b\big)^2\, \mathbb{E}\big[W_2^2\big(\mu_{\tilde{\boldsymbol{Z}}_t'^{-i}}, \mu_{\tilde{\boldsymbol{Z}}_t'}\big)\big] \\ &\leqslant \frac{\big(M_m^b\big)^2}{N(N-1)} \sum_{j:j\neq i} \mathbb{E}\Big[\big|\tilde{Z}_t'^j - \tilde{Z}_t'^i\big|^2\Big] = \frac{2\big(M_m^b\big)^2}{N} \operatorname{Var} P_t^* m.\end{aligned}$$

Finally by (2.70), we have

$$\mathbb{E}\big[\text{(III)}^2\big] \leqslant (M_m^b)^2\, \mathbb{E}\big[W_2^2\big(\mu_{\boldsymbol{Z}_t}, \mu_{\tilde{\boldsymbol{Z}}_t'}\big)\big] \leqslant \frac{\big(M_m^b\big)^2}{N} W_2^2\Big(\big(P_t^N\big)^* m^N, (P_t^* m)^{\otimes N}\Big).$$

Hence

$$\mathbb{E}\big[\delta_t^2\big] = \sum_{i=1}^N \mathbb{E}\big[|\delta_t^i|^2\big] \leqslant C\Big[1 + \operatorname{Var} m_t + W_2^2\Big(\big(P_t^N\big)^* m^N, (P_t^* m)^{\otimes N}\Big)\Big]$$

for some constant $C = C\big(M_m^b, M_{mm}^b\big)$. By Itō's formula the variance $\operatorname{Var} m_t$ satisfies

$$\begin{aligned}\frac{d}{dt} \operatorname{Var} m_t = 2\big(\mathbb{E}\big[\tilde{V}_t^2\big] - \mathbb{E}[\tilde{V}_t]^2\big) + 2\,\mathbb{E}\big[\tilde{X}_t \cdot b(P_t^* m, \tilde{Z}_t)\big] \\ - 2\,\mathbb{E}[\tilde{X}_t] \cdot \mathbb{E}\big[b(P_t^* m, \tilde{Z}_t)\big] + 2d \leqslant C'(\operatorname{Var} m_t + d)\end{aligned}$$

for some $C' = C'(M_z^b, M_m^b)$. Then Grönwall's lemma yields $\operatorname{Var} m_t \leqslant e^{C't} \operatorname{Var} m_0 + (e^{C't} - 1)d$. Upon redefining the constants, we obtain for every $t \geqslant 0$,

$$W_2^2\Big(\big(P_t^N\big)^* m^N, (P_t^* m)^{\otimes N}\Big) \leqslant e^{2M_z t + 1} W_2^2\big(m^N, m^{\otimes N}\big) \\ + Ct \int_0^t e^{2M_z(t-s)} \Big[W_2^2\Big(\big(P_s^N\big)^* m^N, (P_s^* m)^{\otimes N}\Big) + e^{Cs}(\operatorname{Var} m_0 + d)\Big] ds.$$

We then conclude by applying the integral version of Grönwall's lemma. □

2.5.2 From Wasserstein metric to entropy

We study in this section a logarithmic Harnack's inequality for kinetic McKean–Vlasov dynamics and the corresponding particle system. This inequality then implies the regularization from Wasserstein to entropy.

Lemma 2.24 (Log-Harnack inequality for propagation of chaos)**.** *Assume b satisfies (2.64), (2.65) and (2.66), and let $N \geqslant 2$. Then there exist a positive constant C depending on M_x^b, M_v^b, M_m^b and M_{mm}^b such that for every $m \in \mathcal{P}_2(\mathbb{R}^{2d})$ and $m^N \in \mathcal{P}_2(\mathbb{R}^{2dN})$, every measurable function $f^N : \mathbb{R}^{2dN} \to (0, +\infty)$ that is lower bounded away from 0 and upper bounded, and every $T > 0$,*

$$\big(P_T^N \log f^N\big)(m^N) \leqslant \log\big(P_T^{\otimes N} f^N\big)(m) + C\bigg(\frac{1}{(T \wedge 1)^3} + e^{CT}\bigg) W_2^2\big(m^N, m^{\otimes N}\big) \\ + C(e^{CT} - 1)(\operatorname{Var} m + d). \quad (2.71)$$

Consequently,

$$H\Big(\big(P_T^N\big)^* m^N \Big| (P_T^* m)^{\otimes N}\Big) \leqslant C\bigg(\frac{1}{(T \wedge 1)^3} + e^{CT}\bigg) W_2^2\big(m^N, m^{\otimes N}\big) \\ + C(e^{CT} - 1)(\operatorname{Var} m + d). \quad (2.72)$$

Proof. Let us first prove the log-Harnack inequality (2.71) for compactly supported m and m^N.

Constructing a bridge. Fix $T > 0$ and let $(\tilde{X}_t^i, \tilde{V}_t^i)_{i=1}^N$ be N independent duplicates of the solution to (2.67) with the initial condition $\operatorname{Law}(\tilde{X}_0^i, \tilde{V}_0^i) = m$ for $i = 1, \ldots, N$. We denote the N-independent Brownians by $\tilde{W}_t^i$. By enlarging the underlying probability space, we construct random variables $\boldsymbol{X}_0, \boldsymbol{V}_0$ such that

$$\sum_{i=1}^N \mathbb{E}\Big[\big|X_0^i - \tilde{X}_0^i\big|^2 + \big|V_0^i - \tilde{V}_0^i\big|^2\Big] = W_2^2\big(m^N, m^{\otimes N}\big).$$

Define for $i = 1, \ldots, N$ the stochastic processes

$$dX_t^i = V_t^i dt, \quad (2.73)$$

$$dV_t^i = \bigg(b\big(P_t^* m, \tilde{X}_t^i, \tilde{V}_t^i\big) - \frac{V_0^i - \tilde{V}_0^i}{T} + \frac{d}{dt}\big(t(T-t)\big) v^i\bigg) dt + \sqrt{2} d\tilde{W}_t^i, \quad (2.74)$$

where

$$v^i \coloneqq \frac{6}{T^3}\bigg(-\big(X_0^i - \tilde{X}_0^i\big) + \frac{T}{2}\big(V_0^i - \tilde{V}_0^i\big)\bigg). \quad (2.75)$$

The difference processes $\big(X^i_t - \tilde X^i_t, V^i_t - \tilde V^i_t\big)$ satisfy

$$\begin{aligned} d\big(V^i_t - \tilde V^i_t\big) &= -\frac{V^i_0 - \tilde V^i_0}{T}dt + d\big(t(T-t)\big)v^i, \\ d\big(X^i_t - \tilde X^i_t\big) &= \frac{t-T}{T}\big(V^i_0 - \tilde V^i_0\big)dt + t(T-t)v^i dt, \end{aligned}$$

so that

$$V^i_t - \tilde V^i_t = \frac{T-t}{T}\big(V^i_0 - \tilde V^i_0\big) + \frac{6t(T-t)}{T^3}\bigg(-\big(X^i_0 - \tilde X^i_0\big) + \frac{T}{2}\big(V^i_0 - \tilde V^i_0\big)\bigg), \tag{2.76}$$

$$X^i_t - \tilde X^i_t = -\frac{t(t-T)^2}{T}\big(V^i_0 - \tilde V^i_0\big) + \frac{T^3 - 3Tt^2 + 2t^3}{T^3}\big(X^i_0 - \tilde X^i_0\big). \tag{2.77}$$

In particular $X^i_T = \tilde X^i_T$ and $V^i_T = \tilde V^i_T$.

Change of measure. Define

$$\xi^i_t \coloneqq \frac{1}{\sqrt 2}\bigg(b\big(P^*_t m, \tilde X^i_t, \tilde V^i_t\big) - b\big(\mu_{(\boldsymbol{X}_t, \boldsymbol{V}_t)}, X^i_t, V^i_t\big) + \frac{\tilde V^i_0 - V^i_0}{T} + \frac{d}{dt}\big(t(T-t)\big)v^i\bigg)$$

and $\delta b^i_t \coloneqq b\big(P^*_t m, Z^i_t\big) - b\big(\mu_{\boldsymbol{Z}_t}, Z^i_t\big)$. It satisfies

$$\big|\xi^i_t\big| \leqslant C\big|\delta b^i_t\big| + C\bigg(M^b_x + \frac{M^b_v}{T} + \frac{1}{T^2}\bigg)\big(\big|X^i_0 - \tilde X^i_0\big| + T\big|V^i_0 - \tilde V^i_0\big|\big) \tag{2.78}$$

for some universal constant C. In the following C may change from line to line and depend on the constants M^b_x, M^b_v, M^b_m and M^b_{mm}. Set $W^i_\cdot \coloneqq \tilde W^i_\cdot + \int_0^\cdot \xi^i_t dt$ and

$$R_\cdot \coloneqq \exp\bigg[-\sum_{i=1}^N\bigg(\int_0^\cdot \xi^i_t d\tilde W^i_t + \frac12\int_0^\cdot \big|\xi^i_s\big|^2 dt\bigg)\bigg]$$

which is a local martingale. Then (X^i, V^i, W^i) solves (2.68). Since m, m^N are both compactly supported, $\big|X^i_0 - \tilde X^i_0\big|$, $\big|V^i_0 - \tilde V^i_0\big|$ are bounded almost surely. The difference in drift δb^i_t has uniform linear growth in $\boldsymbol{X}_t$, $\boldsymbol{V}_t$, and therefore uniform linear growth in $\tilde{\boldsymbol{X}}_t$, $\tilde{\boldsymbol{V}}_t$. We then apply Lemma B.4 in Appendix B.3 to obtain that $R_\cdot$ is really a martingale. By Girsanov's theorem W^i_t are independent Brownians under the new probability $\mathbb{Q} = R\mathbb{P}$. Since $\boldsymbol{X}_0$, $\boldsymbol{V}_0$, $\tilde{\boldsymbol{X}}_0$, $\tilde{\boldsymbol{V}}_0$ are independent from the Brownian motions we have

$$\sum_{i=1}^N \mathbb{E}^{\mathbb{Q}}\Big[\big|X^i_0 - \tilde X^i_0\big|^2 + \big|V^i_0 - \tilde V^i_0\big|^2\Big] = W_2^2\big(m^N, m^{\otimes N}\big).$$

Hence for measurable functions $f^N : \mathbb{R}^{2dN} \to \mathbb{R}$ that are lower bounded away from 0 and upper bounded, we have

$$\begin{aligned} \big(P^N_T \log f^N\big)(m^N) &= \mathbb{E}\big[R_T \log f^N(\boldsymbol{X}_T, \boldsymbol{V}_T)\big] \\ &\leqslant \mathbb{E}[R_T\log R_T] + \log\mathbb{E}\big[f^N(\boldsymbol{X}_T, \boldsymbol{V}_T)\big] \\ &= \mathbb{E}[R_T\log R_T] + \log\mathbb{E}\big[f^N(\tilde{\boldsymbol{X}}_T, \tilde{\boldsymbol{V}}_T)\big] \\ &= \mathbb{E}[R_T\log R_T] + \log\big(P^{\otimes N}_T f^N\big)(m). \end{aligned}$$

So it remains to bound $\mathbb{E}[R_T \log R_T]$. We observe

$$\mathbb{E}[R_T \log R_T] = \mathbb{E}^{\mathbb{Q}}[\log R_T] = \frac{1}{2}\,\mathbb{E}^{\mathbb{Q}}\bigg[\sum_{i=1}^N \int_0^T |\xi_t^i|^2 dt\bigg]$$

$$\leqslant CT\bigg(M_x^b + \frac{M_v^b}{T} + \frac{1}{T^2}\bigg)^2 \mathbb{E}^{\mathbb{Q}}\Big[\big|X_0^i - \tilde{X}_0^i\big|^2 + T^2\big|V_0^i - \tilde{V}_0^i\big|^2\Big] + C\,\mathbb{E}^{\mathbb{Q}}\bigg[\int_0^T \sum_{i=1}^N \big|\delta b_t^i\big|^2 dt\bigg]$$

$$\leqslant \frac{C(T\vee 1)^3}{(T\wedge 1)^3} W_2^2\big(m^N, m^{\otimes N}\big) + C\,\mathbb{E}^{\mathbb{Q}}\bigg[\int_0^T \sum_{i=1}^N \big|\delta b_t^i\big|^2 dt\bigg].$$

Arguing as in the proof of Proposition 2.23, we have

$$\sum_{i=1}^N \mathbb{E}^{\mathbb{Q}}\Big[\big|\delta b_t^i\big|^2\Big] \leqslant Ce^{Ct}\Big(W_2^2\big(m^N, m^{\otimes N}\big) + \operatorname{Var} m + d\Big),$$

So the log-Harnack inequality (2.71) is proved for compactly supported m^N and m.

Approximation. Now treat general m^N, m of finite second moment, but not necessarily compact supported. Take two sequences $(m_k^N)_{k\in\mathbb{N}}, (m_k)_{k\in\mathbb{N}}$ of compactly supported measures such that $m_k^N \to m^N$ and $m_k \to m$ in respective topologies of $\mathcal{P}_2$. For continuous f^N such that $\log f^N$ is bounded, we have

$$\big(P_t^N \log f^N\big)\big(m_k^N\big) \to \big(P_t^N \log f^N\big)(m^N), \qquad \big(P_t^{\otimes N} f^N\big)(m_k) \to \big(P_t^{\otimes N} f^N\big)(m)$$

by the $\mathcal{P}_2$-continuities of $\big(P_t^N\big)^*$ and P_t^*. So the log-Harnack inequality (2.71) is shown for every continuous f^N which is both lower and upper bounded, and for general m^N and m of finite second moment. For a doubly bounded but not necessarily continuous f^N we take a sequence of continuous and uniformly bounded $(f_k^N)_{k\in\mathbb{N}}$ that converges to f^N in the $\sigma(L^\infty, L^1)$ topology. We have

$$\big(P_t^N \log f_k^N\big)(m^N) \to \big(P_t^N \log f^N\big)(m^N), \qquad \big(P_t^{\otimes N} f_k^N\big)(m) \to \big(P_t^{\otimes N} f^N\big)(m)$$

since both $(P_t^N)^* m^N$ and $P_t^* m$ are absolutely continuous with respect to the Lebesgue measure according to Lemma 2.12. So the desired inequality (2.71) is shown in full generality. Finally, to obtain (2.72) we define another sequence

$$g_k^N \coloneqq \bigg(\frac{\big(P_T^N\big)^* m^N}{(P_T^* m)^{\otimes N}} \wedge k\bigg) \vee \frac{1}{k}$$

for $k \in \mathbb{N}$. We apply the Harnack's inequality (2.71) to g_k^N and take the limit $k \to +\infty$. □

Using the known results on log-Harnack inequalities we can also obtain the regularization in the beginning of the dynamics.

Proposition 2.25. *Assume F satisfies (2.2) and there exist probabilities m_∞, m_∞^N satisfying (2.18), (2.19) respectively and having finite exponential moments. Let m_0 (resp. m_0^N) be the initial value of the mean field dynamics (2.12) (resp. the particle system dynamics (2.13)) of finite second moment. Then there exist a positive constant C depending on M_{mm}^F and M_{mx}^F such that for every $t \in (0,1]$,*

$$H(m_t|m_\infty) \leqslant \frac{C}{t^3} W_2^2(m_0, m_\infty) \quad \textit{(resp. } H\big(m_t^N\big|m_\infty^N\big) \leqslant \frac{C}{t^3} W_2^2\big(m_0^N, m_\infty^N\big)\textit{).} \tag{2.79}$$

Proof. Note that $m_t = P_t^* m_0$ and $m_t^N = \big(P_t^N\big)^* m_0^N$ where P_t and P_t^N are the McKean–Vlasov and the linear semigroup corresponding to the SDEs (2.10), (2.11), respectively. We then apply the log-Harnack inequality for McKean–Vlasov diffusions [193, Proposition 5.1] and obtain

$$H(m_t|m_\infty) \leqslant \frac{C}{t^3} W_2^2(m_t, m_\infty)$$

for $t \in (0,1]$. For the particle system we apply the classical log-Harnack inequality (which corresponds to the case where M_m^b and M_{mm}^b are both equal to 0 in our Lemma 2.24, i.e. no mean field dependence) and obtain

$$H\big(m_t^N\big|m_\infty^N\big) \leqslant \frac{C}{t^3} W_2^2\big(m_t^N, m_\infty^N\big)$$

for $t \in (0,1]$ and it is clear from the computations in Lemma 2.24 that the constant C can be chosen to depend only on M_{mx}^F and M_{mm}^F. □

2.5.3 From entropy to Fisher information

We then adapt Hérau's functional to our setting to obtain the regularization from entropy to Fisher information.

Proposition 2.26. *Assume that F satisfies (2.2) and (2.1), and that there exist probabilities m_∞, m_∞^N satisfying (2.18), (2.19) respectively and having finite exponential moments. Let m_0 (resp. m_0^N) be the initial value of the mean field dynamics (2.12) (resp. the particle system dynamics (2.13)) of finite second moment and finite entropy. Then there exist a positive constant C depending on M_{mm}^F and M_{mx}^F such that for every $t \in (0,1]$,*

$$I(m_t|\hat{m}_t) \leqslant \frac{C}{t^3}\big(\mathcal{F}(m_0) - \mathcal{F}(m_\infty)\big) \quad \textit{(resp. } I\big(m_t^N\big|m_\infty^N\big) \leqslant \frac{C}{t^3} H\big(m_0^N\big|m_\infty^N\big)\big). \tag{2.80}$$

Proof. First derive the bound for the mean field system. We suppose additionally F satisfies (2.32) and $m_0/m_\infty \in \mathcal{A}_+$ without loss of generality, as they can be removed by the approximation argument in the end of the proof of Theorem 2.2. Let a, b, c be positive constants to be determined. Motivated by [221, Theorem A.18], we define Hérau's Lyapunov functional for mean field measures:

$$\begin{aligned}\mathcal{E}(t,m) = F(m) + \frac{1}{2}\int |v^2| m + H(m) + at \int |\nabla_v \log \eta|^2 m \\ + 2bt^2 \int \nabla_x \log \eta \cdot \nabla_v \log \eta m + ct^3 \int |\nabla_x \log \eta|^2 m\end{aligned}$$

where $\eta \coloneqq m/\hat{m}$. From the argument of Theorem 2.2, we know that $\mathcal{E}(t,m_t)$ is well defined and $t \mapsto \mathcal{E}(t,m_t)$ admits derivative satisfying $\frac{d}{dt}\mathcal{E}(t,m_t) \leqslant -Y_t^\top K_t' Y_t$, where K_t' is equal to

$$\begin{pmatrix} 1 - a + 2at - 2\big(M_{mx}^F + M_{mm}^F\big)bt^2 & -2bt^2 & -2at - 4bt - 2M_{mm}^F ct^3 & 0 \\ 0 & 2at & -2M_{mx}^F ct^3 & -4bt^2 \\ 0 & 0 & 2bt^2 - 3ct^2 & 0 \\ 0 & 0 & 0 & 2ct^3 \end{pmatrix}$$

and Y_t is defined by (2.40). We then choose the constants a, b, c depending only on M^F_{mx} and M^F_{mm} such that $ac > b^2$ and $K'_t \succeq 0$ for $t \in [0,1]$. Hence $t \mapsto \mathcal{E}(t, m_t)$ is non-increasing on $[0,1]$ and the Fisher bound follows: for every $t \in (0,1]$,

$$\begin{aligned}
I(m_t|\hat{m}_t) &\leqslant \frac{C}{t^3}\big(\mathcal{E}(t,m_t) - \mathcal{F}(m_t)\big) \\
&\leqslant \frac{C}{t^3}\big(\mathcal{E}(t,m_t) - \mathcal{F}(m_\infty)\big) \\
&\leqslant \frac{C}{t^3}\big(\mathcal{E}(0,m_0) - \mathcal{F}(m_\infty)\big) \\
&= \frac{C}{t^3}\big(\mathcal{F}(m_0) - \mathcal{F}(m_\infty)\big).
\end{aligned}$$

Here, in the second inequality, we use $\mathcal{F}(m_t) \geqslant \mathcal{F}(m_\infty)$ which is a consequence of Lemma 2.9. Note that this inequality relies on the convexity of F.

For the particle system we suppose additionally U^N satisfies (2.47) and m^N_0/m^N_∞ satisfies (2.48) without loss of generality, as they can be removed by the argument in the end of the proof of Lemma 2.19. We define

$$\begin{aligned}
\mathcal{E}^N(t,m^N) = F^N(m^N) &+ \frac{1}{2}\int |\boldsymbol{v}^2| m^N + H(m^N) + at\int \big|\nabla_{\boldsymbol{v}} \log h^N\big|^2 m^N \\
&+ 2bt^2 \int \nabla_{\boldsymbol{x}} \log h^N \cdot \nabla_{\boldsymbol{v}} \log h^N m^N + ct^3 \int \big|\nabla_{\boldsymbol{x}} \log h^N\big|^2 m^N
\end{aligned}$$

where $h^N \coloneqq m^N/m^N_\infty$. By the computations in Lemma 2.19, we have

$$\frac{d}{dt}\mathcal{E}^N\big(t,m^N_t\big) \leqslant -(Y^N_t)^\intercal K''_t Y^N_t,$$

where K''_t is equal to

$$\begin{pmatrix}
1 - a + 2at - 2\big(M^F_{mx} + M^F_{mm}\big)bt^2 & -2bt^2 & -2at - 4bt & 0 \\
 & 2at & -2\big(M^F_{mx} + M^F_{mm}\big)ct^3 & -4bt^2 \\
 & & 2bt^2 - 3ct^2 & 0 \\
 & & & 2ct^3
\end{pmatrix}$$

and Y^N_t is defined by (2.50). We choose again the constants a, b, c depending only on M^F_{mx} and M^F_{mm} such that $ac > b^2$ and $t \mapsto \mathcal{H}^N(t, m^N_t)$ is non-increasing on $[0,1]$. Hence we have for every $t \in (0,1]$,

$$\begin{aligned}
I\big(m^N_t\big|m^N_\infty\big) &\leqslant \frac{C}{t^3}\Big(\mathcal{E}^N\big(t,m^N_t\big) - \mathcal{F}^N\big(m^N_t\big)\Big) \\
&\leqslant \frac{C}{t^3}\Big(\mathcal{E}^N\big(t,m^N_t\big) - \mathcal{F}^N\big(m^N_\infty\big)\Big) \\
&\leqslant \frac{C}{t^3}\Big(\mathcal{E}^N\big(0,m^N_0\big) - \mathcal{F}^N\big(m^N_\infty\big)\Big) \\
&= \frac{C}{t^3}\Big(\mathcal{F}^N\big(m^N_0\big) - \mathcal{F}^N\big(m^N_\infty\big)\Big) \\
&= \frac{C}{t^3} H\big(m^N_0\big|m^N_\infty\big).
\end{aligned}$$

Similarly, we use the fact that $\mathcal{F}^N\big(m^N_t\big) - \mathcal{F}^N\big(m^N_\infty\big) = H\big(m^N_t\big|m^N_\infty\big) \geqslant 0$ to get the second inequality. Here the difference is that the N-particle system is linear and this fact does not rely on the convexity of F. $\square$

2.5.4 Propagation of chaos

Using all the regularization results proved in Sections 2.5.2 and 2.5.3, we can finally give the proof of the main theorem.

Proof of Theorem 2.6. Let m_0 and m_0^N be the respective initial values for the dynamics (2.12), (2.13) and suppose they have finite second moment. The first claim of the theorem (2.22) can be written as two bounds on $W_2^2\big(m_t^N, m_t^{\otimes N}\big)$, the first of which follows directly from the finite-time bound in Proposition 2.23. The second claim (2.23) is nothing but Lemma 2.24. It remains to find some C_2, κ depending only on ρ^x, M_{mx}^F, M_{mm}^F and prove

$$W_2^2\big(m_t^N, m_t^{\otimes N}\big) \leqslant \frac{C_2 N}{(t\wedge 1)^6} W_2^2(m_0, m_\infty) e^{-\kappa t} \\ + \frac{C_2}{(t\wedge 1)^6} W_2^2\big(m_0^N, m_\infty^{\otimes N}\big) e^{-(\kappa - C_2/N)t} + \frac{C_2 d}{\kappa - C_2/N} \tag{2.81}$$

for $t > 0$. Set $t_1 = \frac{t\wedge 1}{2}$ and $t_2 = t\wedge 1$. By the Wasserstein to entropy regularization result in Proposition 2.25, we can find a constant C depending on M_{mx}^F and M_{mm}^F such that

$$H(m_{t_1}|m_\infty) \leqslant \frac{C}{t_1^3} W_2^2(m_0, m_\infty) \quad \text{and} \quad H\big(m_{t_1}^N\big|m_\infty^N\big) \leqslant \frac{C}{t_1^3} W_2^2\big(m_0^N, m_\infty^N\big).$$

In the following C may change from line to line and may depend additionally on the LSI constant ρ. Applying the regularization in Proposition 2.26 to the dynamics with m_{t_1} and $m_{t_1}^N$ as respective initial values and noting that $t_2 - t_1 \leqslant 1$ by definition, we obtain

$$I(m_{t_2}|\hat m_{t_2}) \leqslant \frac{C}{(t_2 - t_1)^3}\big(\mathcal{F}(m_{t_1}) - \mathcal{F}(m_\infty)\big),$$
$$I\big(m_{t_2}^N\big|\hat m_{t_2}^N\big) \leqslant \frac{C}{(t_2 - t_1)^3} H\big(m_{t_1}^N\big|m_\infty^N\big),$$

whereas $\mathcal{F}(m_{t_1}) - \mathcal{F}(m_\infty)$ is bounded by the entropy sandwich in Lemma 2.9:

$$\mathcal{F}(m_{t_1}) - \mathcal{F}(m_\infty) \leqslant C H(m_{t_1}|m_\infty).$$

Consequently, both the measures m_{t_2} and $m_{t_2}^N$ have finite entropy and finite Fisher information, and we can apply respectively Theorems 2.2 and 2.3 to the dynamics with initial values m_{t_2} and $m_{t_2}^N$. We then obtain

$$\mathcal{F}(m_t) - \mathcal{F}(m_\infty) \leqslant \frac{C}{t_1^6} W_2^2(m_0, m_\infty) e^{-\kappa(t - t_2)}$$

and

$$\mathcal{F}^N\big(m_t^N\big) - N\mathcal{F}(m_\infty) \leqslant \frac{C}{t_1^6} W_2^2\big(m_0^N, m_\infty^{\otimes N}\big) e^{-(\kappa - C/N)(t-t_2)} + \frac{Cd}{\kappa - C/N}.$$

Using consecutively the triangle inequality, Talagrand's inequality (2.9) for $m_\infty^{\otimes N}$ and the entropy inequalities in Lemmas 2.9 and 2.10, we have

$$\begin{aligned} W_2^2\big(m_t^N, m_t^{\otimes N}\big) &\leqslant 2W_2^2\big(m_t^N, m_\infty^{\otimes N}\big) + 2NW_2^2(m_t, m_\infty) \\ &\leqslant \frac{2}{\rho}\Big(H\big(m_t^N\big|m_\infty^{\otimes N}\big) + NH(m_t|m_\infty)\Big) \\ &\leqslant \frac{2}{\rho}\Big(\mathcal{F}^N\big(m_t^N\big) - N\mathcal{F}(m_\infty) + N\big(\mathcal{F}(m_t) - \mathcal{F}(m_\infty)\big)\Big) \end{aligned}$$

So the inequality (2.81) is proved by combining the above three inequalities. $\square$

Part II

Log-Sobolev inequalities and singular interactions

Chapter 3

Logarithmic Sobolev inequalities for non-equilibrium steady states

Abstract. We consider two methods to establish log-Sobolev inequalities for the invariant measure of a diffusion process when its density is not explicit and the curvature is not positive everywhere. In the first approach, based on the Holley–Stroock and Aida–Shigekawa perturbation arguments [*J. Stat. Phys.*, 46(5-6):1159–1194, 1987; *J. Funct. Anal.*, 126(2):448–475, 1994], the control on the (non-explicit) perturbation is obtained by stochastic control methods, following the comparison technique introduced by Conforti [*Ann. Appl. Probab.*, 33(6A):4608–4644, 2023]. The second method combines the Wasserstein-2 contraction method, used in [*Ann. Henri Lebesgue*, 6:941–973, 2023] to prove a Poincaré inequality in some non-equilibrium cases, with Wang's hypercontractivity results [*Ann. Probab.*, 37(4):1587–1604, 2009].

Based on joint work with Pierre Monmarché.

3.1 Introduction

3.1.1 Overview

A probability measure μ on $\mathbb{R}^d$ is said to satisfy a logarithmic Sobolev inequality (LSI) with constant $C_{\mathrm{LS}} > 0$ if for all smooth and compactly supported function f on $\mathbb{R}^d$ with $\int f^2 \,\mathrm{d}\mu = 1$, we have

$$\int_{\mathbb{R}^d} f^2 \ln f^2 \,\mathrm{d}\mu \leqslant 2C_{\mathrm{LS}} \int_{\mathbb{R}^d} |\nabla f|^2 \,\mathrm{d}\mu\,.$$

It is related to the long-time convergence of diffusion processes and concentration inequalities, see [12] and references therein for general considerations on this topic. The main question addressed here is to establish such an LSI in cases where μ has no explicit density but is defined as the invariant measure of a diffusion process $(Z_t)_{t\geqslant 0}$ satisfying

$$\mathrm{d}Z_t = b(Z_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t\,, \tag{3.1}$$

where $b \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$, σ is a constant matrix and B is a d-dimensional Brownian motion. Among other applications, this is motivated by non-equilibrium statistical physics models, such as [122, 123, 67]. In this literature, these non-explicit invariant measures are referred to as Non-Equilibrium Steady States (NESS).

Many criteria to establish LSI are known, but most of them require an explicit expression for μ (for instance in order to use bounded or Lipschitz perturbation arguments) or the process (3.1) to be reversible with respect to μ (for instance the Lyapunov-based results of [10, 42]). In fact, denoting by

$$\mathcal{L} = b \cdot \nabla + \Sigma : \nabla^2$$

the generator of (3.1) (where $\Sigma = \sigma\sigma^T/2$ and $\Sigma : \nabla^2 = \sum_{i,j} \Sigma_{ij}\partial_{z_i}\partial_{z_j}$), some arguments based on reversibility (such as those of [10, 42]) may sometimes be extended to non-reversible cases when the dual $\mathcal{L}^*$ of $\mathcal{L}$ in $L^2(\mu)$ is known. Since

$$\mathcal{L}^* = (2\Sigma\nabla \ln \mu - b) \cdot \nabla + \Sigma : \nabla^2 \,,$$

knowing $\mathcal{L}^*$ requires an explicit expression of μ. A notable exception is the use of Bakry–Émery curvature conditions: if there exists $\rho > 0$ such that, for all x, $y \in \mathbb{R}^d$,

$$\big(b(x) - b(y)\big) \cdot (x - y) \leqslant -\rho |x - y|^2 \,, \tag{3.2}$$

then the diffusion (3.1) admits a unique invariant measure that satisfies an LSI with constant $|\Sigma|/\rho$, see [41, 165], even when the process is neither reversible nor elliptic. However, such a contraction condition is very restrictive. If (3.2) holds only for x, y outside some compact set, we can decompose $b = b_0 + b_1$ where b_0 satisfies a similar condition on the whole $\mathbb{R}^d$ and b_1 is compactly supported. Then, we know that the invariant measure μ_0 of the process with generator $\mathcal{L}_0 = b_0 \cdot \nabla + \Sigma : \nabla^2$ satisfies an LSI, but to the best of our knowledge it is not known how to transfer the result to μ in this general case.

In this work, we will consider two cases:

- In the first one, μ is a perturbation of an explicit measure μ_0, invariant for $\mathcal{L}_0 = b_0 \cdot \nabla + \Sigma : \nabla^2$ for some b_0. Our method relies on the bounded perturbation result of Holley and Stroock [113] and the Lipschitz perturbation result of Aida and Shigekawa [1]. In other words, the key point is to prove that $\ln(\mu/\mu_0)$ is the sum of a bounded function and a Lipschitz function. This is done by seeing this quantity as the long-time limit of the solution of a (parabolic) Hamilton–Jacobi–Bellman (HJB) equation and using a stochastic control representation for the solution together with a coupling argument, following the method introduced by Conforti in [61]. This approach is applied to the elliptic case and to a non-elliptic kinetic case.

- In the second one, we consider the high-diffusivity elliptic framework of [168], namely (3.2) holds for every y once x lies outside some compact set and $\sigma = \bar{\sigma}\mathrm{Id}$ where $\bar{\sigma} > 0$ is large enough. Under these conditions, on the one hand, it is proven in [168] that μ satisfies a Poincaré inequality, using the large-time contraction of the Wasserstein-2 distance along the diffusion semi-group. On the other hand, as established by Wang in [228] (in the reversible case but we will see that the proof applies without any change in the non-reversible case), the semi-group is hypercontractive, which implies a so-called defective LSI (which is well known in the reversible case and turns out to be true in general). The Poincaré inequality together with the defective LSI is equivalent to the LSI.

In fact, thanks to the powerful [223, Corollary 1.2], the defective LSI alone is already equivalent to a tight LSI for irreducible diffusion processes. This has been used in subsequent works [226, 225, 121] for both elliptic and kinetic processes. However this argument is non-constructive and thus does not provide explicit constants, similarly to the tightening argument based on weak Poincaré inequalities in [198, Proposition 1.3]. This is in our contrast to our approach, as illustrated in Chapter 4 which is based on the present work.

The rest of this work is organized as follows. The results are stated in the remainder of Section 3.1. The results based on perturbation, Theorems 3.1 and 3.6, are proven in Section 3.2. Section 3.3 is devoted to the proofs for the defective LSI. A coupling construction for the kinetic Langevin process, used in the proof of Theorem 3.6, is postponed to Appendix 3.4.

3.1.2 Perturbation approach: the elliptic case

In the elliptic case where $\Sigma = \mathrm{Id}$, we get the following, proven in Section 3.2.1.

Theorem 3.1. *Assume that $\Sigma = \mathrm{Id}$ and $b = b_0 + b_1$ for some b_0, $b_1 \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$ with bounded derivatives such that the generator $\mathcal{L}_0 = b_0 \cdot \nabla + \Delta$ admits a unique C^2 and positive invariant probability density μ_0 satisfying an LSI with constant $C_0 > 0$. Write $\tilde{b} := 2\nabla \ln \mu_0 - b$ and $\varphi := -\nabla \cdot b_1 + b_1 \cdot \nabla \ln \mu_0$. Assume that there exist L, R, M^φ, $L^\varphi \geqslant 0$ and $\rho > 0$ such that $\varphi = \varphi_1 + \varphi_2$ with φ_1 being M^φ-bounded and φ_2 being L^φ-Lipschitz, and for all x, $y \in \mathbb{R}^d$,*

$$\big(\tilde{b}(x) - \tilde{b}(y)\big) \cdot (x - y) \leqslant \begin{cases} -\rho|x-y|^2 & \textit{if } |x-y| \geqslant R, \\ L|x-y|^2 & \textit{otherwise.} \end{cases} \tag{3.3}$$

Finally, assume that the law of Z_t solving (3.1) converges weakly for all initial condition as $t \to \infty$ to a unique invariant measure μ on $\mathbb{R}^d$. Then μ satisfies an LSI with constant $C_{\mathrm{LS}} = C_{\mathrm{LS}}(C_0, \rho, L, R, d; M^\varphi, L^\varphi)$.

Notice that, when $\Sigma = \mathrm{Id}$, the carré du champ operator $\Gamma(f) = \frac{1}{2}\mathcal{L}(f^2) - f\mathcal{L}f$ is equal to $|\nabla f|^2$ and is the same for the dual operator $\mathcal{L}^*$. In particular, the LSI is equivalent to the constant-rate decay of the relative entropy,

$$\forall \nu \ll \mu\,, \qquad \mathcal{H}(P_t^*\nu|\mu) \leqslant e^{-t/C_{\mathrm{LS}}}\mathcal{H}(\nu|\mu)\,, \tag{3.4}$$

where $P_t = \exp(t\mathcal{L})$ is the semi-group generated by $\mathcal{L}$, and the relative entropy $\mathcal{H}$ is defined by $\mathcal{H}(\nu|\mu) = \int \ln(\mathrm{d}\nu/\mathrm{d}\mu)\,\mathrm{d}\nu$ (see e.g. [12, Theorem 5.2.1]; reversibility is not used in the proof).

Example 3.2. Consider on $\mathbb{R}^2$ the SDE

$$\mathrm{d}X_t = \big(f(|X_t|)X_t^\perp - X_t - \nabla V(X_t)\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,,$$

where $(u,v)^\perp = (v,-u)$, $f \in \mathcal{C}^1(\mathbb{R}_+, \mathbb{R})$ and $V \in \mathcal{C}^2(\mathbb{R}^2)$. We can decompose the drift $b = b_0 + b_1$ as

$$b_0(x) = x^\perp f(|x|) - x\,, \qquad b_1(x) = -\nabla V(x)\,, \tag{3.5}$$

or, alternatively,

$$b_0(x) = -\nabla V(x) - x\,, \qquad b_1(x) = f(|x|)x^{\perp}\,. \tag{3.6}$$

In the first case (3.5), the invariant measure of $b_0 \cdot \nabla + \Delta$ is the standard Gaussian measure $\mu_0(x) = (2\pi)^{-d/2}\exp(-|x|^2/2)$ and, using the notations of Theorem 3.1,

$$\tilde{b}(x) = -f(|x|)x^{\perp} - x + \nabla V(x)\,, \qquad \varphi(x) = \Delta V(x) + \nabla V(x)\cdot x\,.$$

For instance, if $f(|x|)$ is constant for $|x|$ large enough, then the rotating part does not intervene in the condition (3.3) outside some compact set, which means that this condition is satisfied as soon as $(x-y)\cdot\big(\nabla V(x) - \nabla V(y)\big) \leqslant \eta|x-y|^2$ outside some compact set for some $\eta < 1$. In this situation, ergodicity for (3.1) is easily shown using Harris Theorem. Then Theorem 3.1 applies as soon as φ is Lipschitz, which is for instance the case if V is compactly supported (which implies the previous condition). Notice that here we do not use the fact that the perturbative term $b_1 = -\nabla V$ is a gradient.

If, alternatively, we use the decomposition (3.6), then μ_0 is the probability density proportional to $\exp\big(-|x|^2/2 - V(x)\big)$ and

$$\tilde{b}(x) = -f(|x|)x^{\perp} - x - \nabla V(x)\,, \qquad \varphi(x) = f(|x|)x^{\perp}\cdot\nabla V(x)\,.$$

Then, Theorem 3.1 applies for instance if f is compactly supported and $(x-y)\cdot\big(\nabla V(x) - \nabla V(y)\big) \geqslant -\eta|x-y|^2$ for some $\eta < 1$ outside some compact set.

Example 3.3. Let us check how Theorem 3.1 reads in the classical reversible case, namely taking

$$b_0(x) = -\nabla U(x)\,, \qquad b_1(x) = -\nabla W(x)$$

for some U, $W \in \mathcal{C}^2(\mathbb{R}^d)$, so that $\mu_0 \propto e^{-U}$, $\mu \propto e^{-U-W}$ and, with the notations of Theorem 3.1,

$$\tilde{b}(x) = -\nabla U(x) + \nabla W(x)\,, \qquad \varphi(x) = \Delta W(x) + \nabla U(x)\cdot\nabla W(x)\,.$$

Hence, the conditions in Theorem 3.1 does not seem to be similar to those of classical perturbation results in the reversible case. However, for instance, if we take $U(x) = |x|^2/2$ outside some compact set, then to get that $x \mapsto \nabla U(x)\cdot\nabla W(x)$ is Lipschitz one will typically require ∇W to be bounded, in which case the LSI for μ follows from [1].

Example 3.4. We now consider a non-linear McKean-Vlasov equation on $\mathbb{R}^d$:

$$\partial_t \mu_t = \nabla\cdot\big((\nabla V + \lambda b_{\mu_t})\mu_t + \nabla\mu_t\big) \tag{3.7}$$

where $x \mapsto \mu_t(x)$ is a probability density on $\mathbb{R}^d$ (which we identify with the corresponding probability measure), $V \in \mathcal{C}^2(\mathbb{R}^d)$ is a confining potential, $\lambda \in \mathbb{R}$ encodes the non-linearity amplitude, $\nabla\cdot$ stands for the divergence operator and, for all probability measure ν, $b_\nu \in \mathcal{C}^1(\mathbb{R}^d)$. If μ_* is a stationary measure for (3.7), it is the invariant measure of the diffusion process with generator $\mathcal{L}_{\mu_*}$ where

$$\mathcal{L}_\mu = -(\nabla V + \lambda b_\mu)\cdot\nabla + \Delta\,.$$

Among other examples of interest where, for a given μ, b_μ is not the gradient of some potential, we can mention the competition models considered in [156], where $x = (x_1, x_2) \in \mathbb{R}^{2p}$ and

$$b_\mu(x_1,x_2) = \begin{pmatrix} \int_{\mathbb{R}^p} \nabla_{x_1} K(x_1,y_2)\mu(y_1,y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ -\int_{\mathbb{R}^p} \nabla_{x_2} K(y_1,x_2)\mu(y_1,y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \end{pmatrix} \tag{3.8}$$

for some $K \in \mathcal{C}^2(\mathbb{R}^{2p})$. In other words, the population is divided in two types of individuals, the first type (resp. second) tends to maximize (resp. minimize) its value of K averaged with respect to the population distribution of the other type.

In order to get an LSI for μ_*, for instance, it is straightforward to check that the assumptions of Theorem 3.1 are met (with $\mu_0 \propto e^{-V}$) under the following condition:

Assumption 3.5. The potential V is strictly convex outside a compact, and its hessian is bounded. There exists L', $C' > 0$ such that, for all μ, b_μ is L'-Lipschitz and for all $x \in \mathbb{R}^d$,

$$b_\mu(x) \leqslant C' \frac{1 + \int_{\mathbb{R}^d} |y|\mu(\mathrm{d}y)}{1 + |x|} . \tag{3.9}$$

In particular, the condition (3.9) is used to get that, for a given μ with finite expectation, φ is bounded (as $|\nabla \ln \mu_0(x)| = |\nabla V(x)| \leqslant C(1+|x|)$ for some $C > 0$). The condition that $\nabla^2 V$ is bounded can be lifted if (3.9) is replaced by a stronger decay of $b_\mu(x)$.

For instance, in the case (3.8), the condition (3.9) holds when $|\nabla_x K(x,y)| \leqslant C/(1 + |x-y|^2)$ for some constant $C > 0$ (and similarly for $\nabla_y K$). Indeed, then, considering the first coordinate of (3.8) (the second one being similar), we bound

$$\begin{aligned} \left| \int_{\mathbb{R}^p} \nabla_{x_1} K(x_1,y_2)\mu(y_1,y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \right| &\leqslant C \int_{\mathbb{R}^p} \frac{1}{1+|x-y_2|^2}\mu(y_1,y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &\leqslant \frac{C}{1+|x|} + C\mathbb{P}_\mu[|Y_2| \geqslant |x| - \sqrt{|x|}] \end{aligned}$$

and the Markov inequality concludes. The fact that a stationary solution of (3.7) has a finite expectation is implied by Assumption 3.5 for λ small enough, as e.g. $w(x) = |x|^2$ is a Lyapunov function for (3.7).

Contrary to the linear case, obtaining an LSI for μ_* is however not sufficient to get the exponential convergence of the solution of (3.7) toward μ_*, as, even under Assumption 3.5, several stationary solutions may exist [110]. However, this is sufficient to conclude in the weak interaction regime (i.e. when λ is small enough), provided the interaction drift is Lipschitz in terms of the non-linearity:

$$\exists B > 0 \text{ s.t. } \forall \nu, \nu', \qquad \|b_\nu - b_{\nu'}\|_\infty \leqslant B\mathcal{W}_2(\nu,\nu') , \tag{3.10}$$

where the $\mathcal{W}_2$-Wasserstein distance between two probability measures ν, ν' is defined as

$$\mathcal{W}_2(\nu,\nu') = \inf_{\pi \in \mathcal{C}(\nu,\nu')} \left(\int_{(\mathbb{R}^d)^2} |x-x'|^2 \pi(\mathrm{d}x,\mathrm{d}x') \right)^{1/2} ,$$

with $\mathcal{C}(\nu,\nu')$ the set of probability measures on $(\mathbb{R}^d)^2$ with marginal ν and ν'.

Indeed, in that case, by a classical computation,

$$
\begin{aligned}
\partial_t \mathcal{H}(\mu_t|\mu_*) &= \int_{\mathbb{R}^d} \partial_t (\ln \mu_t)\mu_t + \ln \frac{\mu_t}{\mu_*}\partial_t \mu_t \\
&= \partial_t \int_{\mathbb{R}^d} \mu_t + \int_{\mathbb{R}^d} \mathcal{L}_{\mu_t}\left(\ln \frac{\mu_t}{\mu_*}\right)\mu_t \\
&= \int_{\mathbb{R}^d} \mathcal{L}_{\mu_*}\left(\ln \frac{\mu_t}{\mu_*}\right)\mu_t + \int_{\mathbb{R}^d} (\mathcal{L}_{\mu_t} - \mathcal{L}_{\mu_*})\left(\ln \frac{\mu_t}{\mu_*}\right)\mu_t \\
&= -\int_{\mathbb{R}^d} \left|\nabla \ln \frac{\mu_t}{\mu_*}\right|^2 \mu_t + \lambda \int_{\mathbb{R}^d} \nabla \ln \frac{\mu_t}{\mu_*} \cdot (b_{\mu_*} - b_{\mu_t})\mu_t \\
&\leqslant -\frac{1}{2}\int_{\mathbb{R}^d} \left|\nabla \ln \frac{\mu_t}{\mu_*}\right|^2 \mu_t + \frac{1}{2}B^2\lambda^2 \mathcal{W}_2^2(\mu_t, \mu_*)\,,
\end{aligned}
$$

where we used Cauchy–Schwarz and (3.10). Now, the LSI satisfied by μ_* implies the Talagrand inequality

$$
\mathcal{W}_2^2(\mu_t, \mu_*) \leqslant C\mathcal{H}(\mu_t|\mu_*)
$$

(where C is the LSI constant of μ_*), see [180]. As a consequence, using that the LSI constant of μ_* is uniformly bounded over small values of λ (since Theorem 3.1 can be applied with the same constants L^φ and M^φ for all values of $\lambda \in [0, \lambda_0]$ for any $\lambda_0 > 0$), we get that

$$
\partial_t \mathcal{H}(\mu_t|\mu_*) \leqslant -\varepsilon \mathcal{H}(\mu_t|\mu_*)
$$

for some $\varepsilon > 0$ for λ small enough. As a conclusion, we obtain that μ_* is the unique stationary solution of (3.7) and globally attractive.

Notice that (3.10) holds for the model (3.8) as soon as $\nabla^2 K$ is bounded, since in that case,

$$
\begin{aligned}
&\left| \int_{\mathbb{R}^p} \nabla_{x_1} K(x_1, y_2)\nu(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 - \int_{\mathbb{R}^p} \nabla_{x_1} K(x_1, y_2)\nu'(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \right| \\
&\qquad\qquad\qquad \leqslant \|\nabla^2 K\|_\infty \int_{(\mathbb{R}^p)^2} |y_2 - y_2'|\pi(\mathrm{d}y_1\,\mathrm{d}y_2, \mathrm{d}y_1\,\mathrm{d}y_1')\,,
\end{aligned}
$$

where π is any coupling of ν and ν', so that conclusion follows by Cauchy–Schwarz and taking the infimum over all couplings (the second coordinate of $b_\nu - b_{\nu'}$ being treated similarly).

3.1.3 Perturbation approach: the kinetic case

We consider in this section a non-equilibrium Langevin diffusion $Z = (X, V)$ on $\mathbb{R}^d \times \mathbb{R}^d$ solving

$$
\left\{ \begin{array}{rcl} \mathrm{d}X_t &=& V_t\,\mathrm{d}t \\ \mathrm{d}V_t &=& -\nabla U(X_t)\,\mathrm{d}t + G(X_t, V_t)\,\mathrm{d}t - \gamma V_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t \end{array} \right. \tag{3.11}
$$

for some $\gamma > 0$, $U \in \mathcal{C}^2(\mathbb{R}^d)$, $G \in \mathcal{C}^1(\mathbb{R}^{2d}, \mathbb{R}^d)$, as studied in [122, 169]. The particular case where G depends only on V correspond to non-linear friction models, see e.g. [160, 137].

Contrary to the elliptic case, now, the LSI is not equivalent to the entropy decay (3.4). However, from the LSI, the (hypocoercive) decay of the entropy along $(P_t)_{t\geqslant 0}$ (rather than $(P_t^*)_{t\geqslant 0}$) can be obtained applying Theorems 9 and 10 of [166] even without the explicit knowledge of the invariant measure.

Theorem 3.6. *Assume that* e^{-U} *is integrable and that the probability measure with density proportional to* e^{-U} *satisfies an LSI with constant* C_0*. Let*

$$\varphi(x,v) = -\nabla_v G(x,v) + G(x,v)\cdot v\,.$$

Assume that φ *is* L^φ*-Lipschitz and the drift writes*

$$-\nabla U(x) + G(x,-v) = -Kx + g(x,v)$$

for a positive-definite matrix K *whose smallest eigenvalue is* $k>0$*, and a function* $g:\mathbb{R}^{2d}\to\mathbb{R}$ *satisfying*

$$|g(x,v)-g(x',v')| \leqslant \begin{cases} L_1|z-z'| & \text{if } |x-x'|+|v-v'|\leqslant R,\\ L_2|z-z'| & \text{otherwise,}\end{cases}$$

where $|z-z'| = \sqrt{|x-x'|^2+|v-v'|^2}$ *is the Euclidean distance, for some constants* R*,* L_1*,* $L_2\geqslant 0$*. If additionally* $19\max(1,\gamma)L_2\leqslant\min(1,k)$ *and the law of* (X_t,V_t) *solving* (3.11) *converges weakly for all initial condition as* $t\to\infty$ *to a unique invariant measure* μ *on* $\mathbb{R}^{2d}$*, then* μ *satisfies an LSI with a constant* $C_{\mathrm{LS}} = C_{\mathrm{LS}}(C_0,K,L_1,L_2,R,\gamma;L^\varphi)$*.*

The proof of this result is given in Section 3.2.2.

Remark 3.7. The assumptions of the kinetic perturbation theorem seem to be more restrictive than the elliptic one. First, the drift in the kinetic case must be the sum of a positive linear transform plus a perturbation term whose oscillation "grows slowly enough" compared to the linear term, while in the elliptic case it only needs to satisfy a weak convexity condition. This is because our proof is based on W_1-contraction of diffusion processes and in the kinetic case such contraction is harder to establish (e.g. compare Theorem 3.20 to [83]). Second, the function φ in the kinetic case must be Lipschitz while in the elliptic case it can be the sum of a Lipschitz and a bounded function, due to the fact that the coupling of Theorem 3.20 does not allow us to obtain total variation bounds (dual to bounded functions) as is done for the elliptic case.

3.1.4 Defective LSI approach

A measure $\mu\in\mathcal{P}(\mathbb{R}^d)$ is said to satisfy a defective log-Sobolev inequality if for all $f\geqslant 0$ with $\int f\,\mathrm{d}\mu = 1$,

$$\int_{\mathbb{R}^d} f\ln f\,\mathrm{d}\mu \leqslant A\int_{\mathbb{R}^d}\frac{|\nabla f|^2}{f}\,\mathrm{d}\mu + B\,, \tag{3.12}$$

for some constants A, $B\geqslant 0$. From [12, Proposition 5.1.3], such a defective LSI, together with a Poincaré inequality

$$\forall f\in L^2(\mu),\qquad \int_{\mathbb{R}^d}\left(f-\int_{\mathbb{R}^d} f\,\mathrm{d}\mu\right)^2\mathrm{d}\mu \;\leqslant\; C\int_{\mathbb{R}^d}|\nabla f|^2\,\mathrm{d}\mu \tag{3.13}$$

for some constant $C > 0$, implies an LSI for μ with constant $A' = A + C(B+2)/4$ (i.e. (3.12) but with A replaced by A' and B replaced by 0).

In some non-reversible elliptic cases, a Poincaré inequality has been established in [168] (see Proposition 3.12 below). To improve this result into an LSI, it is thus enough to obtain a defective LSI.

In the following for $\alpha, \beta \geqslant 1$ we write $\|f\|_\alpha = (\int |f|^\alpha \,\mathrm{d}\mu)^{1/\alpha}$ and

$$\|P_t\|_{\alpha\to\beta} = \sup\{\|P_t f\|_\beta \ : \ f \in L^\alpha(\mu), \|f\|_\alpha = 1\}\,,$$

where μ is the invariant measure of the semi-group P_t considered. The semi-group is said to be hypercontractive if there exist $t_0 > 0$, $\alpha < \beta$ such that $\|P_{t_0}\|_{\alpha\to\beta} < \infty$. In that case without loss of generality we can assume that $\alpha = 1$, as the following easily follows from Hölder's inequality (the proof is given in Section 3.3.1 for completeness):

Lemma 3.8. *For all $\alpha, \gamma > 1$,*

$$\|P_t\|_{1\to\alpha} \leqslant \|P_t\|^{\gamma\alpha-1}_{\alpha\to(\gamma\alpha-1)/(\gamma-1)}\,.$$

In the reversible settings, it is well-known that hypercontractivity implies a defective LSI, see [12]. We show that it is also true in the non-reversible case. For simplicity, we only consider the case where the diffusion matrix Σ is constant, since this is anyway the case in [168]. In the reversible case, the proof relies on [12, Proposition 5.2.6], whose proof requires reversibility. In the non-reversible case, we replace this result by the following (proven in Section 3.3):

Proposition 3.9. *Let $(P_t)_{t\geqslant 0}$ be a diffusion semi-group with invariant measure μ and generator $b\cdot\nabla + \Sigma : \nabla^2$ where Σ is a constant diffusion matrix and b satisfies the one-sided Lipschitz condition*

$$\forall x, y \in \mathbb{R}^d, \qquad \big(b(x) - b(y)\big)\cdot(x-y) \leqslant L|x-y|^2\,. \tag{3.14}$$

Then, for all $f \geqslant 0$ with $\int_{\mathbb{R}^d} f\,\mathrm{d}\mu = 1$, all $\alpha > 0$ and all $t \geqslant 0$,

$$\int_{\mathbb{R}^d} f\ln f\,\mathrm{d}\mu \leqslant \frac{\alpha+1}{\alpha}\ln\|P_t\|_{1\to1+\alpha} + |\Sigma|\frac{e^{2Lt}-1}{2L}\int_{\mathbb{R}^d}\frac{|\nabla f|^2}{f}\,\mathrm{d}\mu\,.$$

In the remaining of this section, we focus on the following elliptic case.

Assumption 3.10. The semi-group $(P_t)_{t\geqslant 0}$, whose generator reads $b\cdot\nabla + \sigma\Delta$ for $\sigma > 0$, admits an invariant measure μ and there exist L, ρ, $R > 0$ such that

$$\forall x, y \in \mathbb{R}^d, \qquad \big(b(x) - b(y)\big)\cdot(x-y) \leqslant \begin{cases} -\rho|x-y|^2 & \text{if } |x| \geqslant R,\\ L|x-y|^2 & \text{otherwise.}\end{cases} \tag{3.15}$$

Note that this assumption is different from (3.3), which we imposed for the perturbation result in the elliptic case. Under this assumption, hypercontractivity follows from the Harnack inequality established by Wang in [228] (originally stated in the reversible case but the proof, recalled in Section 3.3.2, is unchanged in the non-reversible one). More specifically, we get the following.

Proposition 3.11. *Let $\beta > \alpha > 1$. Under Assumption 3.10, set*

$$t_0 = \frac{2\beta}{\sigma^2 \rho(\alpha - 1)} .$$

Then, for all $t > t_0$,

$$\|P_t\|_{\alpha \to \beta} \leqslant \big(1 + 4d + 2(L+\rho)R^2\big) \exp\left(\frac{\beta L R t}{2\sigma^2(\alpha - 1)} + \frac{1}{8} \max\left(\frac{1+4d}{t/t_0 - 1}, 2\rho R^2 \right) \right) .$$

Combining Propositions 3.9 and 3.11 gives a defective LSI (see Corollary 3.19). As a conclusion, we recall the following result from [168, Theorems 1 and 2].

Proposition 3.12. *Under Assumption 3.10, assume furthemore that*

$$\sigma \geqslant \sigma_0 \coloneqq (2L+\rho) \frac{(2L+\rho/2)R_*^2 + 2\sup\{-x \cdot b(x),\ |x| \leqslant R_*\}}{\rho d} , \tag{3.16}$$

where $R_* = R(2 + 2L/\rho)^{1/d}$. Then μ satisfies the Poincaré inequality (3.13) with constant

$$C = \frac{4\sigma}{\rho} \left(1 + \frac{\alpha(2L+\rho)R_*^2}{4d\sigma} \right) . \tag{3.17}$$

Thanks to [12, Proposition 5.1.3], the defective LSI of Corollary 3.19 and the Poincaré inequality of Proposition 3.12 yields the following.

Corollary 3.13. *Under Assumption 3.10, provided furthemore (3.16), μ satisfies an LSI with constant $C_{\mathrm{LS}} = A + C(B+2)/4$ where A, B, C are respectively given in (3.27), (3.28), (3.17).*

As in Section 3.1.2, in the present elliptic case, the LSI is equivalent to the entropy decay (3.4).

3.2 Proofs

3.2.1 The elliptic case

Before proving the theorem, let us first show a key lemma on the value function of stochastic optimal control problems.

Lemma 3.14. *Let $U \subset \mathbb{R}^d$. Under the conditions of Theorem 3.1, consider the stochastic optimal control problem,*

$$V(T,x) = \sup_{\nu} \sup_{\alpha:\alpha_t \in U} \mathbb{E}\left[\int_0^T \big(\varphi(X_t^{\alpha,x}) - |\alpha_t|^2 \big) \,\mathrm{d}t \right] ,$$

where $\nu = \big(\Omega, F, (\mathcal{F}_\cdot), \mathbb{P}, (B_\cdot)\big)$ stands for a filter probability space with the usual conditions and an $(\mathcal{F}_\cdot)$-Brownian motion, α is an $\mathbb{R}^d$-valued progressively measurable process such that $\int_0^T \mathbb{E}\big[|\alpha_t|^m\big]\,\mathrm{d}t$ is finite for every $m \in \mathbb{N}$, and $X^{\alpha,x}$ solves

$$X_0^{\alpha,x} = x , \qquad \mathrm{d}X_t^{\alpha,x} = \big(\tilde{b}(X_t^{\alpha,x}) + 2\alpha_t\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t .$$

Then there exists $C' > 0$, depending only on ρ, L, R, such that for every x, $y \in \mathbb{R}^d$, every $T > 0$ and every $t \in (0,T]$, we have

$$|V(T,x) - V(T,y)| \leqslant 2M^\varphi t + C' \left(\frac{2M^\varphi}{t} + L^\varphi \right) |x - y| .$$

Proof. The method has been demonstrated in [61] and we give a proof for the sake of completeness. Fix $\varepsilon > 0$ and $(T, x) \in (0, +\infty) \times \mathbb{R}^d$. Take an ε-optimal control $(\nu^\varepsilon, \alpha^\varepsilon)$ such that

$$V(T, x) \leqslant \mathbb{E}\biggl[\int_0^T \bigl(\varphi(X_t) - |\alpha_t^\varepsilon|^2\bigr)\,\mathrm{d}t\biggr] + \varepsilon\,,$$

where we denote $X = X^{\alpha^\varepsilon, x}$. Construct the process Y solving

$$Y_0 = y\,, \qquad \mathrm{d}Y_t = \bigl(\tilde{b}(Y_t) + 2\alpha_t^\varepsilon\bigr)\,\mathrm{d}t + \sqrt{2}(1 - 2e_t e_t^{\mathrm{T}})\,\mathrm{d}B_t\,,$$

until $\tau \coloneqq \inf\{t : X_t = Y_t\}$ and $Y_t = X_t$ henceforth, where e_t is defined by

$$e_t = \begin{cases} \frac{X_t - Y_t}{|X_t - Y_t|} & \text{if } X_t \neq Y_t, \\ (1, 0, \ldots, 0)^{\mathrm{T}} & \text{otherwise.} \end{cases}$$

Then the difference process $\delta X_t \coloneqq X_t - Y_t$ solves

$$\mathrm{d}\delta X_t = \bigl(\tilde{b}(X_t) - \tilde{b}(Y_t)\bigr)\,\mathrm{d}t + 2\sqrt{2} e_t e_t^{\mathrm{T}}\,\mathrm{d}B_t\,.$$

Thanks to the weak convexity condition (3.3), there exist $C \geqslant 0$ and $\kappa > 0$ such that

$$\begin{aligned} \mathbb{E}\bigl[|X_t - Y_t|\bigr] &\leqslant Ce^{-\kappa t}|x - y|\,, \\ \mathbb{P}\bigl[X_t \neq Y_t\bigr] &\leqslant \frac{Ce^{-\kappa t}}{t}|x - y|\,, \end{aligned}$$

where the first inequality is due to Eberle [83] and the second to the sticky coupling [86, Theorem 3]. By the definition of V, we have

$$V(T, y) \geqslant \mathbb{E}\biggl[\int_0^T \bigl(\varphi(Y_t) - |\alpha_t^\varepsilon|^2\bigr)\,\mathrm{d}t\biggr]\,.$$

Hence by subtracting the expressions for $V(T, x)$ and $V(T, y)$, we obtain

$$\begin{aligned} V(T, x) - V(T, y) &\leqslant \biggl(\int_0^t + \int_t^T\biggr)\,\mathbb{E}\bigl[\varphi_1(X_s) - \varphi_1(Y_s)\bigr]\,\mathrm{d}s + \int_0^T \mathbb{E}\bigl[\varphi_2(X_s) - \varphi_2(Y_s)\bigr]\,\mathrm{d}s \\ &\leqslant 2M^\varphi t + \frac{C}{\kappa}\biggl(\frac{2M^\varphi}{t} + L^\varphi\biggr)|x - y| + \varepsilon\,. \end{aligned}$$

Taking $\varepsilon \to 0$ gives the desired upper bound for $V(T, x) - V(T, y)$. The lower bound follows by exchanging x and y. □

Now we present the proof for the perturbation result in the elliptic case. We will use the notion of viscosity solution and we refer readers to [64, Section 8] for its definition.

Proof of Theorem 3.1. Under the conditions of Theorem 3.1, consider $m_t = \mathrm{Law}(Z_t)$ where Z solves (3.1) with initial distribution $m_0 = \mu_0$. Then m_t, μ_0 solve respectively

$$\begin{aligned} \partial_t m_t &= -\nabla \cdot (b m_t) + \Delta m_t\,, \\ 0 = \partial_t \mu_0 &= -\nabla \cdot (b_0 \mu_0) + \Delta \mu_0 \end{aligned}$$

where the first equation holds in the sense of distributions a priori. By approximation arguments, we can show that $(t,x) \mapsto m_t(x)$ is continuous and a viscosity solution to the first equation. Define the relative density $h_t = m_t/\mu_0$. Then, it is a viscosity solution to

$$\partial_t h_t = \Delta h_t + \tilde{b}_t \cdot \nabla h_t + \varphi h_t\,, \tag{3.18}$$

where $\varphi = -\nabla \cdot b_1 + b_1 \cdot \nabla \ln \mu_0$ and $\tilde{b} = -b + 2\nabla \ln \mu_0$. Notice that the value of h_t can be given by the Feynman–Kac formula

$$h_t(x) = \mathbb{E}\left[\exp\left(\int_0^t \varphi(X_s^{t,x})\,\mathrm{d}s\right) h_0(X_t^{t,x})\right],$$

where $X^{t,x}$ solves

$$X_0^{t,x} = x, \qquad dX_s^{t,x} = \tilde{b}_{t-s}(X_s^{t,x})\,\mathrm{d}s + \sqrt{2}dB_t \quad \text{for } s \in [0,t].$$

Suppose additionally that φ is bounded and Lipschitz continuous. Then applying synchronous coupling to the Feynman–Kac formula above, we obtain a constant $M > 0$ such that

$$M^{-1} \leqslant h(t,x) \leqslant M \qquad \text{and} \qquad |h(t,x) - h(s,y)| \leqslant M\big(|t-s|^{1/2} + |x-y|\big)$$

for every t, $s \in [0,T]$ and every x, $y \in \mathbb{R}^d$. Taking the logarithm $u_t \coloneqq \ln h_t$ and using the fact that $h \mapsto \ln h$ is a strictly increasing and C^2 mapping, we obtain that u_t is a bounded and uniformly continuous viscosity solution to the HJB equation,

$$\partial_t u_t = \Delta u_t + |\nabla u_t|^2 + \tilde{b}_t \cdot \nabla u_t + \varphi\,. \tag{3.19}$$

The rest of the proof then amounts to linking the HJB equation to the stochastic optimal control problem considered in Lemma 3.14.

For $N \in \mathbb{N}$, consider the approximative HJB equation,

$$\partial_t u_t^N = \Delta u_t^N + \sup_{\alpha : |\alpha| \leqslant N} \{2\alpha \cdot \nabla u_t^N - |\alpha|^2\} + \tilde{b} \cdot \nabla u^N + \varphi\,, \tag{3.20}$$

and the associated control problem,

$$V^N(T,x) = \sup_{\nu} \sup_{\alpha : |\alpha_t| \leqslant N} \mathbb{E}\left[\int_0^T \big(\varphi(X_t^{\alpha,x}) - |\alpha_t|^2\big)\,\mathrm{d}t\right], \tag{3.21}$$

where ν, α satisfy the conditions in the statement of Lemma 3.14. By Theorem IV.7.1 and the results in Sections V.3 and V.9 of [91], the value function V^N defined by (3.21) is a bounded and uniformly continuous viscosity solution to (3.20). Applying Lemma 3.14 to the approximative problem (3.21), we obtain a constant $C' > 0$ such that

$$\begin{aligned} |V^N(t,x) - V^N(t,y)| &\leqslant C' \|\varphi\|_{\mathrm{Lip}} |x-y|\,, \\ |V^N(T,x) - V^N(T,y)| &\leqslant 2M^\varphi t + C'\left(\frac{2M^\varphi}{t} + L^\varphi\right)|x-y| \end{aligned}$$

for every $t \in (0,T]$ and every x, $y \in \mathbb{R}^d$. Hence if $w \in C^{1,2}\big([0,T) \times \mathbb{R}^d\big)$ is such that $V^N - w$ attains a local maximum or a local minimum at $(t,x) \in [0,T) \times \mathbb{R}^d$, then $|\nabla w(t,x)| \leqslant C'\|\varphi\|_{\mathrm{Lip}}$ by the first inequality above. This implies that V^N is actually a viscosity solution to the original (3.19) for $N \geqslant C'\|\varphi\|_{\mathrm{Lip}}$. Since both u and V^N are bounded and uniformly continuous on $[0,T] \times \mathbb{R}^d$, we can apply the parabolic comparison for viscosity solutions on the whole space [72, Theorem 1] to obtain $V^N(T,x) = u_T(x)$ for N sufficiently large. Therefore, for every $T > 0$, every $t \in (0,T]$ and every x, $y \in \mathbb{R}^d$, we have

$$|u_T(x) - u_T(y)| \leqslant 2M^\varphi t + C'\left(\frac{2M^\varphi}{t} + L^\varphi\right)|x-y|\,. \tag{3.22}$$

Now we remove the additional assumption on φ and take a sequence of $\varphi^n = \varphi^{n,1} + \varphi^{n,2}$ such that each of φ^n is bounded and Lipschitz continuous, $\|\varphi^{n,1}\|_\infty \leqslant M^\varphi$, $\|\varphi^{n,2}\|_{\mathrm{Lip}} \leqslant L^\varphi$ for all $n \in \mathbb{N}$, and $\varphi^n \to \varphi$ locally uniformly. For each n, consider the equation

$$\partial_t h_t^n = \Delta h_t^n + \tilde{b} \cdot \nabla h_t^n + \varphi^n h_t^n$$

and let h^n be the solution given by the Feynman–Kac formula with the initial condition $h_0^n = 1$. Taking the limit $n \to +\infty$ in the Feynman–Kac formulas and using the dominated convergence theorem, we obtain that $h_T^n \to h_T$ pointwise. Yet, each $u_T^n \coloneqq \ln h_T^n$ satisfies the bound (3.22) when u is replaced by u^n. So taking the limit, we obtain that (3.22) still holds without the additional assumption on φ.

Denote the Gaussian kernel in $\mathbb{R}^d$ by $g^\varepsilon = (2\pi\varepsilon)^{-d/2}\exp(-|x|^2/2\varepsilon)$. We decompose u_T in the following way:

$$u_T = u_T \star g^\varepsilon + (u_T - u_T \star g^\varepsilon).$$

Thanks to (3.22), we find that the first term is uniformly Lipschitz, and the second term is uniformly bounded. Then we apply successively the Holley–Stroock and Aida–Shigekawa perturbation lemmas [113, 1], and obtain that the flow of measures

$$(m_T)_{T\geqslant 1} = (\mu_0 \exp u_T)_{T\geqslant 1}$$

satisfies a uniform log-Sobolev inequality (see [40, Theorem 2.7] for an explicit constant for Aida–Shigekawa). Noticing that the LSI is stable under the weak convergence of measures, we take the limit $T \to +\infty$ and conclude. □

Remark 3.15. We exploit the properties of viscosity solution to the HJB equation (3.19) instead of classical solution, contrary to what is done by Conforti [61]. The main reason for this is that we wish to be able to treat the kinetic, therefore degenerate elliptic, case in the same framework, for which the existence of classical solution, despite the system's hypoellipticity, is lacking in classical literatures of stochastic optimal control to our knowledge.

3.2.2 The kinetic case

As in the previous section, we first establish a lemma on the kinetic stochastic optimal control problem.

Lemma 3.16. *Let $U \subset \mathbb{R}^d$. Under the conditions of Theorem 3.6, consider the stochastic optimal control problem,*

$$V(T,z) = \sup_{\nu} \sup_{\alpha:\alpha_t\in U} \mathbb{E}\left[\int_0^T \left(\varphi(Z_t^{\alpha,z}) - \gamma|\alpha_t|^2\right)\mathrm{d}t\right],$$

where $\nu = \big(\Omega, F, (\mathcal{F}.), \mathbb{P}, (B.)\big)$ stands for a filter probability space with the usual conditions and an $(\mathcal{F}.)$-Brownian motion, α is an $\mathbb{R}^d$-valued progressively measurable process such that $\int_0^T \mathbb{E}\big[|\alpha_t|^m\big]\,\mathrm{d}t$ is finite for every $m \in \mathbb{N}$, and $Z^{\alpha,z} = (X^{\alpha,z}, Y^{\alpha,z})$ solves

$$Z_0^{\alpha,x} = z\,, \quad \begin{cases} \mathrm{d}X_t^{\alpha,z} = V_t^{\alpha,z}\,\mathrm{d}t\,, \\ \mathrm{d}V_t^{\alpha,z} = \big(-\gamma V_t^{\alpha,z} - \nabla U(X_t^{\alpha,z}) + G(X_t^{\alpha,z}, -V_t^{\alpha,z}) + 2\gamma\alpha_t\big)\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,. \end{cases}$$

Then there exists $C' > 0$, depending only on K, L_1, L_2, R, γ, such that for every z, $z' \in \mathbb{R}^d$, we have

$$|V(T,z) - V(T,z')| \leqslant C' L^\varphi |z - z'|\,.$$

Proof. Fix $\varepsilon > 0$ and $(T,z) \in (0,+\infty) \times \mathbb{R}^{2d}$. Take an ε-optimal control $(\nu^\varepsilon, \alpha^\varepsilon)$ such that

$$V(T,z) \leqslant \mathbb{E}\bigg[\int_0^T \big(\varphi(Z_t) - \gamma|\alpha_t^\varepsilon|^2\big)\,\mathrm{d}t\bigg] + \varepsilon\,,$$

where we denote $(X,V) = Z = Z^{\alpha^\varepsilon,z}$. Using γt as the new time variable and $\gamma^{-1}X$ as the new space variable, and noticing that $-\nabla U(x) + G(x,-v) = -Kx + g(x,v)$, we can apply Theorem 3.20 in the appendix to construct the processes $Z_n' = (X_n', Y_n')$ solving

$$Z_{n,0}' = z'\,, \qquad \begin{cases} \mathrm{d}X_{n,t}' = V_{n.t}'\,\mathrm{d}t\,, \\ \mathrm{d}V_{n,t}' = \big(-\gamma V_{n,t}' - KX_{n,t}' + g(Z_{n,t}') + 2\gamma\alpha_t\big)\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_{n,t}'\,, \end{cases}$$

where B_n' are Brownian motions, and there exist constant $C_1 \geqslant 1$, $\kappa > 0$ such that

$$\limsup_{n\to+\infty} \mathbb{E}\big[|Z_t - Z_{n,t}'|\big] \leqslant C_1 e^{-\kappa t}\rho(z,z') \quad \text{for } t \geqslant 0.$$

By the definition of V, we have

$$V(T,z') \geqslant \mathbb{E}\bigg[\int_0^T \big(\varphi(Z_{n,t}') - \gamma|\alpha_t^\varepsilon|^2\big)\,\mathrm{d}t\bigg]\,.$$

Hence by subtracting the expressions for $V(T,z)$ and $V(T,z')$, we obtain

$$V(T,z) - V(T,z') \leqslant \int_0^T \mathbb{E}\big[\varphi(Z_s) - \varphi(Z_{n,s}')\big]\,\mathrm{d}s + \varepsilon \leqslant L^\varphi \int_0^T \mathbb{E}\big[|Z_s - Z_{n,s}'|\big]\,\mathrm{d}s + \varepsilon\,,$$

By Fatou's lemma we have

$$\begin{aligned} \liminf_{n\to+\infty} \int_0^T \mathbb{E}\big[|Z_s - Z_{n,s}'|\big]\,\mathrm{d}s \leqslant \int_0^T \liminf_{n\to+\infty} \mathbb{E}\big[|Z_s - Z_{n,s}'|\big]\,\mathrm{d}s \leqslant \int_0^T C_1 e^{-\kappa t}\rho(z,z')\,\mathrm{d}s \\ < \frac{C_1}{\kappa}\rho(z,z')\,. \end{aligned}$$

So taking $\varepsilon \to 0$ and exchanging x and y, we obtain

$$|V(T,z) - V(T,z')| \leqslant \frac{C_1 L^\varphi}{\kappa}\rho(z,z')\,.$$

Setting the interpolation $z_s = (1-s)z + sz'$, by the previous inequality we have

$$\begin{aligned}|V(T,z) - V(T,z')| &= \sum_{i=0}^{N-1} \bigl|V\bigl(T, z_{(i+1)/N}\bigr) - V\bigl(T, z_{i/N}\bigr)\bigr| \\ &\leqslant \frac{C_1 L^\varphi}{\kappa} \sum_{i=0}^{N-1} \rho\bigl(z_{(i+1)/N}, z_{i/N}\bigr) \\ &= \frac{C_1 L^\varphi |z-z'|}{\kappa} \frac{1}{N} \sum_{i=0}^{N-1} \frac{\rho\bigl(z_{(i+1)/N}, z_{i/N}\bigr)}{\bigl|z_{(i+1)/N} - z_{i/N}\bigr|}\,.\end{aligned}$$

As the uniform convergence $\limsup_{z\to z'} \rho(z,z')/|z-z'| \leqslant C_2$ holds, we have

$$\limsup_{N\to+\infty} \frac{1}{N} \sum_{i=0}^{N-1} \frac{\rho\bigl(z_{(i+1)/N}, z_{i/N}\bigr)}{\bigl|z_{(i+1)/N} - z_{i/N}\bigr|} \leqslant C_2\,.$$

Thus taking the limit $N \to +\infty$ in the inequality above concludes the proof. □

Proof of Theorem 3.6. Under the conditions of Theorem 3.6, consider $m_t = \mathrm{Law}(Z_t)$ where Z solves (3.11) with the initial distribution

$$m_0(\mathrm{d}x\,\mathrm{d}v) = \mu_0(\mathrm{d}x\,\mathrm{d}v) \propto \exp\biggl(-U(x) - \frac{1}{2}|v|^2\biggr)\,\mathrm{d}x\,\mathrm{d}v\,,$$

which is the unique invariant measure of the diffusion (3.11) when $G = 0$. By the tensorization property, μ_0 satisfies an LSI with constant $\max(1, C_0)$. The measures m_t, μ_0 solve respectively

$$\begin{aligned}\partial_t m_t &= \gamma\Delta_v m_t + \nabla_v \cdot \bigl[m_t\bigl(\gamma v + \nabla U - G(x,v)\bigr)\bigr] - v\cdot\nabla_x m_t\,,\\ 0 = \partial_t \mu_0 &= \gamma\Delta_v \mu_0 + \nabla_v \cdot [\mu_0(\gamma v + \nabla U)] - v\cdot\nabla_x \mu_0\,,\end{aligned}$$

where the first equation holds in the sense of viscosity. Define the relative density $h_t = m_t/\mu_0$. Then, it is a viscosity solution to

$$\partial_t h_t = \gamma\Delta_v h_t + \bigl(-\gamma v + \nabla U(x) - G(x,v)\bigr)\cdot\nabla_v h - v\cdot\nabla_x h + \varphi h\,, \tag{3.23}$$

where $\varphi = -\nabla_v G(x,v) + G(x,v)\cdot v$. Taking the logarithm $u_t \coloneqq \ln h_t$ and using the fact that $h \mapsto \ln h$ is a strictly increasing and C^2 mapping, we obtain that u_t is a viscosity solution to the kinetic HJB equation,

$$\partial_t u_t = \gamma\Delta_v u_t + \gamma|\nabla_v u_t|^2 + \bigl(-\gamma v + \nabla U(x) - G(x,v)\bigr)\cdot\nabla_v u - v\cdot\nabla_x u + \varphi\,. \tag{3.24}$$

Now, on the formal level the kinetic HJB equation (3.24) is related to the optimal control problem considered in Lemma 3.16: if the domain of control in the lemma is unrestricted, i.e. $U = \mathbb{R}^d$, then we expect to have

$$u_T(x,v) = V(T,x,-v)\,.$$

We then argue as in the proof of Theorem 3.1 (suppose φ is regular enough, then restrict the domain of control, finally approximate for general φ) to validate this claim. Then by Lemma 3.16, for every z, $z' \in \mathbb{R}^{2d}$, we have

$$|u_T(z) - u_T(z')| \leqslant C' L^\varphi |z-z'|\,.$$

We conclude as in the end of the proof of Theorem 3.1. □

3.3 Defective log-Sobolev inequality

3.3.1 From Hypercontractivity to defective LSI

Proof of Proposition 3.9. From [165, Theorem 1], the one-sided Lipschitz condition (3.14) implies that $|\nabla P_t f| \leqslant e^{tL} P_t |\nabla f|$ for all $t \geqslant 0$ for all f. Then, classically, for $f \geqslant 0$ with $\int f \, \mathrm{d}\mu = 1$,

$$\begin{aligned}
\int_{\mathbb{R}^d} P_t f \ln P_t f \, \mathrm{d}\mu &= \int_{\mathbb{R}^d} f \ln f \, \mathrm{d}\mu - \int_0^t \int_{\mathbb{R}^d} \frac{|\Sigma^{1/2} \nabla P_s f|^2}{P_s f} \, \mathrm{d}\mu \, \mathrm{d}s \\
&\geqslant \int_{\mathbb{R}^d} f \ln f \, \mathrm{d}\mu - |\Sigma| \int_0^t \int_{\mathbb{R}^d} \frac{e^{2sL} (P_s |\nabla f|)^2}{P_s f} \, \mathrm{d}\mu \, \mathrm{d}s \\
&\geqslant \int_{\mathbb{R}^d} f \ln f \, \mathrm{d}\mu - |\Sigma| \int_0^t \int_{\mathbb{R}^d} \frac{e^{2sL} |\nabla f|^2}{f} \, \mathrm{d}\mu \, \mathrm{d}s \,,
\end{aligned}$$

where $\big(P_s(|\nabla f|)\big)^2 \leqslant P_s(|\nabla f|^2/f) P_s(f)$ (by Cauchy–Schwarz) and the invariance of μ by P_s were used in the last inequality. In other words,

$$\begin{aligned}
\int_{\mathbb{R}^d} f \ln f \, \mathrm{d}\mu &\leqslant \int_{\mathbb{R}^d} P_t f \ln P_t f \, \mathrm{d}\mu + |\Sigma| \frac{e^{2Lt} - 1}{2L} \int_{\mathbb{R}^d} \frac{|\nabla f|^2}{f} \, \mathrm{d}\mu \\
&= \frac{1}{\alpha} \int_{\mathbb{R}^d} P_t f \ln (P_t f)^\alpha \, \mathrm{d}\mu + |\Sigma| \frac{e^{2Lt} - 1}{2L} \int_{\mathbb{R}^d} \frac{|\nabla f|^2}{f} \, \mathrm{d}\mu \,,
\end{aligned}$$

for every $\alpha > 0$. By Jensen's inequality applied to the probability measure $P_t f \mu$,

$$\int_{\mathbb{R}^d} P_t f \ln (P_t f)^\alpha \, \mathrm{d}\mu \leqslant \ln \int_{\mathbb{R}^d} (P_t f)^{1+\alpha} \, \mathrm{d}\mu \leqslant (1+\alpha) \ln \|P_t\|_{1 \to 1+\alpha} \,,$$

since $\|P_t f\|_1 = 1$. □

Proof of Lemma 3.8. Using Hölder's inequality, for $f \geqslant 0$ with $\int f \, \mathrm{d}\mu = 1$ (so that $\|P_t f\|_1 = 1$ since μ is invariant by P_t),

$$\begin{aligned}
\|P_t f\|_\alpha^\alpha &\leqslant \|P_t f\|_1^{1/\gamma} \|P_t f\|_{(\gamma\alpha-1)/(\gamma-1)}^{(\gamma\alpha-1)/\gamma} \\
&\leqslant \|P_t\|_{\alpha \to (\gamma\alpha-1)/(\gamma-1)}^{(\gamma\alpha-1)/\gamma} \|P_t f\|_\alpha^{(\gamma\alpha-1)/\gamma} .
\end{aligned}$$

Dividing by $\|P_t f\|_\alpha^{(\gamma\alpha-1)/\gamma}$ concludes. □

3.3.2 Hypercontractivity in the elliptic case

Next, we recall (here in a non-reversible settings – which doesn't change the proof – and only in the flat space ; also with explicit constants) the Harnack inequality of [228].

Proposition 3.17. *Assume that there exists $K > 0$ such that*

$$\forall x, y \in \mathbb{R}^d, \qquad (x-y) \cdot \big(b(x) - b(y)\big) \leqslant K |x-y| \,. \tag{3.25}$$

Then, for all $t \geqslant 0$, all $x, y \in \mathbb{R}^d$ and all $\alpha > 1$,

$$\big(P_t f(y)\big)^\alpha \leqslant (P_t f^\alpha)(x) \exp\left(\frac{\alpha}{2\sigma^2(\alpha-1)} \left(K^2 t + \frac{|x-y|^2}{t} \right) \right) .$$

Notice that, in particular, Assumption 3.10 implies (3.25) with $K = LR$.

Proof. For two initial conditions $x \neq y$ and a final time $T > 0$, let X, Y solve

$$\begin{aligned} X_0 &= x\,, \qquad \mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t\,, \\ Y_0 &= y\,, \qquad \mathrm{d}Y_t = b(Y_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t + \xi e_t\,\mathrm{d}t\,, \end{aligned}$$

where $e_t = (X_t - Y_t)/|X_t - Y_t|$ for $t < \tau := \inf\{s \geqslant 0,\ X_s = Y_s\}$ and $e_t = 0$ for $t \geqslant \tau$ (so that in particular $X_t = Y_t$ for $t \geqslant \tau$) and $\xi = K + |x-y|/T$.

Since the norm is $\mathcal{C}^2$ outside the origin, we can apply Itō's formula up to time τ to get, for $t < \tau$,

$$\begin{aligned} \mathrm{d}|X_t - Y_t| &= e_t \cdot \big(b(X_t) - b(Y_t)\big)\,\mathrm{d}t - \xi_t\,\mathrm{d}t \\ &\leqslant -\frac{|x-y|}{T}\,\mathrm{d}t\,. \end{aligned}$$

This implies that $\tau \leqslant T$, and thus $X_T = Y_T$. By Girsanov's theorem,

$$P_T f(y) = \mathbb{E}[f(Y_T)R]\,, \quad \text{with} \quad R = e^{\frac{\xi}{\sigma}\int_0^\tau e_t\cdot\mathrm{d}B_t - \frac{\xi^2}{2\sigma^2}\tau}\,,$$

so that, by Hölder's inequality, for $f \geqslant 0$ and $\alpha > 1$, using that $Y_T = X_T$,

$$\big(P_T f(y)\big)^\alpha \leqslant (P_T f^\alpha)(x)\big(\mathbb{E}\,R^{\alpha/(\alpha-1)}\big)^{\alpha-1}$$

with

$$\big(\mathbb{E}\,R^{\alpha/(\alpha-1)}\big)^{\alpha-1} \leqslant \exp\left(\frac{\alpha}{4\sigma^2(\alpha-1)}\xi^2 T\right) \leqslant \exp\left(\frac{\alpha}{2\sigma^2(\alpha-1)}\left(K^2 T + \frac{|x-y|^2}{T}\right)\right).$$

□

Lemma 3.18. *Under Assumption 3.10,*

$$\int_{\mathbb{R}^d} e^{\delta|x-y|^2}\mu(\mathrm{d}x)\mu(\mathrm{d}y) \leqslant \big(1 + 4d + (2L + 8\delta)R^2\big)\exp\left(\delta\max\left(\frac{1+4d}{2(\rho - 4\delta)}, R^2\right)\right),$$

for all $\delta \in (0, \rho/4)$.

Proof. Let $\mathcal{L}_2$ be the generator of two independent diffusion processes satisfying (3.1), so that $\mu \otimes \mu$ is invariant by $\mathcal{L}_2$. For $\delta \in (0, \rho/4)$, consider the function $V(x,y) = e^{\delta|x-y|^2}$. Then

$$\begin{aligned} \frac{\mathcal{L}_2 V(x,y)}{V(x,y)} &= 2\delta(x-y)\cdot\big(b(x) - b(y)\big) + 4\delta d + 8\delta^2|x-y|^2 \\ &\leqslant \begin{cases} 4\delta d + (8\delta^2 - 2\delta\rho)|x-y|^2 & \text{if } |x-y| \geqslant R \\ 4\delta d + (2\delta L + 8\delta^2)R^2 & \text{otherwise} \end{cases} \\ &\leqslant -\delta\mathbb{1}_{|x-y|\geqslant R_*} + \delta\big(4d + (2L + 8\delta)R^2\big)\mathbb{1}_{|x-y|<R_*} \end{aligned}$$

with

$$R_*^2 = \max\left(\frac{1+4d}{2(\rho-4\delta)}, R^2\right).$$

Hence

$$\begin{aligned}\mathcal{L}_2 V(x,y) &\leqslant -\delta V(x,y) + \delta\big(1+4d+(2L+8\delta)R^2\big)\mathbb{1}_{|x-y|\leqslant R_*} V(x,y)\\ &\leqslant -\delta V(x,y) + \delta\big(1+4d+(2L+8\delta)R^2\big)e^{\delta R_*^2}\,.\end{aligned}$$

Integrating with respect to $\mu\otimes\mu$, the left hand side vanishes and we get

$$\int_{\mathbb{R}^d} V(x,y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) \leqslant \big(1+4d+(2L+8\delta)R^2\big)e^{\delta R_*^2}\,,$$

as announced. □

Proof of Proposition 3.11. Let $f\geqslant 0$ be such that $\mu(f^\alpha)=1$. By Proposition 3.17 (with $K=LR$), for any $y\in\mathbb{R}^d$,

$$\begin{aligned}1 &= \int_{\mathbb{R}^d} P_t(f^\alpha)(x)\mu(\mathrm{d}x)\\ &\geqslant \big(P_t f(y)\big)^\alpha \int_{\mathbb{R}^d}\exp\left(\frac{-\alpha}{2\sigma^2(\alpha-1)}\left(K^2 t + \frac{|x-y|^2}{t}\right)\right)\mu(\mathrm{d}x)\,.\end{aligned}$$

As a consequence, for $\beta>\alpha$,

$$\begin{aligned}\int_{\mathbb{R}^d}\big(P_t f(y)\big)^\beta\mu(\mathrm{d}y) &\leqslant \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\exp\left(\frac{-\alpha}{2\sigma^2(\alpha-1)}\left(Kt+\frac{|x-y|^2}{t}\right)\right)\mu(\mathrm{d}x)\right]^{-\beta/\alpha}\mu(\mathrm{d}y)\\ &\leqslant \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\left(\frac{\beta}{2\sigma^2(\alpha-1)}\left(Kt+\frac{|x-y|^2}{t}\right)\right)\mu(\mathrm{d}x)\mu(\mathrm{d}y)\,.\end{aligned}$$

Conclusion follows from (3.18) and using that $t/t_0\leqslant 1$. □

To conclude, gathering Propositions 3.9 and 3.11, we get the following:

Corollary 3.19. *Assume* (3.15) *for some* L, $R\geqslant 0$ *and* $\rho>0$. *Then for all* $f\geqslant 0$ *with* $\int_{\mathbb{R}^d} f\,\mathrm{d}\mu = 1$, *we have*

$$\int_{\mathbb{R}^d} f\ln f \leqslant A\int_{\mathbb{R}^d}\frac{|\nabla f|^2}{f}\,\mathrm{d}\mu + B \tag{3.26}$$

with

$$A = \frac{\sigma^2}{2L}\left(\exp\left(\frac{24L}{\sigma^2\rho}\right)-1\right), \tag{3.27}$$

$$B = 6\ln\big(1+4d+2(L+\rho)R^2\big) + \frac{108LR}{\sigma^4\rho} + \frac{3}{4}\max\big(1+4d, 2\rho R^2\big) \tag{3.28}$$

(taking for A *the limit as* $L\to 0$ *of this expression if* $L=0$*).*

Proof. For simplicity, take $\alpha=1$ in Proposition 3.9, $\alpha=\gamma=2$ in Lemma 3.8 and $t=2t_0$ in Proposition 3.11, we end up with

$$\begin{aligned}\int_{\mathbb{R}^d} f\ln f &\leqslant 2\ln\|P_t\|_{1\to 2} + \sigma^2\frac{e^{2Lt}-1}{2L}\int_{\mathbb{R}^d}\frac{|\nabla f|^2}{f}\,\mathrm{d}\mu\\ &\leqslant 6\ln\|P_t\|_{2\to 3} + \sigma^2\frac{e^{2Lt}-1}{2L}\int_{\mathbb{R}^d}\frac{|\nabla f|^2}{f}\,\mathrm{d}\mu\,,\end{aligned}$$

and conclusion follows the expression given in Proposition 3.11. □

3.4 Reflection coupling for kinetic diffusions

Theorem 3.20 (Coupling by reflection for kinetic diffusions). *Let $\big(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geqslant 0}, \mathbb{P}\big)$ be a filtered probability space satisfying the usual conditions. Let X, V be $\mathbb{R}^d$-valued continuous and adapted processes, and α be an $\mathbb{R}^d$-valued progressively measurable process solving*

$$\begin{aligned} \mathrm{d}X_t &= V_t \,\mathrm{d}t\,, \\ \mathrm{d}V_t &= \alpha_t \,\mathrm{d}t + \big({-V_t} - KX_t + g(X_t, V_t)\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \end{aligned} \tag{3.29}$$

for $t \geqslant 0$, where K is a $d\times d$ symmetric and positive-definite matrix, $g : \mathbb{R}^{2d} \to \mathbb{R}^d$ is a Lipschitz continuous function, and $(B_t)_{t\geqslant 0}$ is an $(\mathcal{F}_\cdot)$-Brownian motion in d dimensions. Let X'_0, V'_0 be $\mathbb{R}^d$-valued and $\mathcal{F}_0$-measurable random variables. Denote by k the smallest eigenvalue of K. Suppose that $\int_0^T \mathbb{E}\big[|\alpha_t|^2\big]\,\mathrm{d}t$ is finite for every $T > 0$, and X_0, V_0, X'_0, V'_0 are all square-integrable. If there exist nonnegative constants R, L_1, L_2 such that for every z, $z' \in \mathbb{R}^{2d}$, we have

$$\big|g(z) - g(z')\big| \leqslant \begin{cases} L_1|z-z'| & \text{if } |x-x'| + |v-v'| \leqslant R, \\ L_2|z-z'| & \text{otherwise,} \end{cases}$$

with $L_2 < \frac{1}{19}\min(1,k)$ and $L_2 \leqslant L_1$, then upon enlarging the probability space, we can construct a sequence of continuous and adapted processes X'_n, V'_n such that

1. *their initial values are given by X'_0, V'_0, that is, $X'_{n,0} = X'_0$ and $V'_{n,0} = V'_0$;*
2. *they solve*

$$\begin{aligned} \mathrm{d}X'_{n,t} &= V'_{n,t} \,\mathrm{d}t\,, \\ \mathrm{d}V'_{n,t} &= \alpha_t \,\mathrm{d}t + \big({-V'_{n,t}} - KX'_{n,t} + b(X'_{n,t}, V'_{n,t})\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_{n,t}\,, \end{aligned} \tag{3.30}$$

   for $(\mathcal{F}_\cdot)$-Brownians $(B'_{n,t})_{t\geqslant 0}$;
3. *and finally, there exists constants C_1, $C_2 \geqslant 1$, $\kappa > 0$ and a continuous function of quadratic growth $\rho : \mathbb{R}^{2d}\times\mathbb{R}^{2d} \to \mathbb{R}$, all explicitly expressible by K, R, L_1, L_2, such that*

$$\limsup_{n\to+\infty} \mathbb{E}\big[|Z_t - Z'_{n,t}|\big] \leqslant C_1 e^{-\kappa t}\, \mathbb{E}\big[\rho(Z_0, Z'_0)\big] \quad \text{for } t \geqslant 0, \tag{3.31}$$

   and uniformly in z', we have

$$\limsup_{z\to z'} \frac{\rho(z,z')}{|z-z'|} \leqslant C_2\,.$$

Remark 3.21. We develop a translation-invariant version of the additive metric constructed in [84] under which the difference processes, $\delta X = X - X'$, $\delta V = V - V'$, are contractive in average. Our assumptions are an improvement over [128, Theorem 2.16] although we do not elaborate on mean field dependence. Also, we can recover the contraction in W_1 distance from the transport cost ρ by a limiting procedure (as is done in the proof of Lemma 3.16). So our approach can be used to achieve the W_1-contraction of [206, Theorem 5] with simpler calculations.

Proof. For x, x', v, $v' \in \mathbb{R}^d$, introduce the variables $q = x + v$, $q' = x' + v'$ and denote $\delta x = x - x'$, $\delta v = v - v'$, $\delta q = q - q'$. Define

$$r(z, z') = \theta|\delta x| + |\delta q| = \theta|x - x'| + |q - q'|$$

where

$$\theta \coloneqq 2\max\bigl(|K| + L_1, 1\bigr).$$

Denote $r_0 = (\theta + 1)R$.

Reflection-synchronous coupling. Fix an $n \in \mathbb{N}$. Let us construct the desired processes $Z'_n = (X'_n, V'_n)$. Find Lipschitz-continuous rc_n, $\mathrm{sc}_n : \mathbb{R}^{2d} \times \mathbb{R}^{2d} \to \mathbb{R}$ satisfying $\mathrm{rc}_n^2 + \mathrm{sc}_n^2 = 1$ and

$$\mathrm{rc}_n(z, z') = \begin{cases} 0 & \text{if } r(z, z') \geqslant r_0 + n^{-1} \text{ or } |\delta q| \leqslant n^{-1}, \\ 1 & \text{if } r(z, z') \leqslant r_0 \text{ and } |\delta q| \geqslant 2n^{-1}. \end{cases}$$

Upon enlarging the filtered probability space, we can also find another $(\mathcal{F}_\cdot)$-Brownian motion B'' that is independent from B. [1] Let $Z'_n = (X'_n, V'_n)$ solve

$$\begin{aligned} \mathrm{d}X'_{n,t} &= V'_{n,t}\,\mathrm{d}t\,, \\ \mathrm{d}V'_{n,t} &= \alpha_t\,\mathrm{d}t + \bigl(-V'_{n,t} - KX'_{n,t} + g(X'_{n,t}, V'_{n,t})\bigr)\,\mathrm{d}t \\ &\qquad + \mathrm{rc}_n(Z_t, Z'_{n,t})(1 - 2e_{n,t}e_{n,t}^{\mathrm{T}})\sqrt{2}\,\mathrm{d}B^{\mathrm{rc}}_{n,t} + \mathrm{sc}_n(Z_t, Z'_{n,t})\sqrt{2}\,\mathrm{d}B^{\mathrm{sc}}_{n,t}\,, \end{aligned}$$

with initial value $Z'_{n,0} = (X'_0, V'_0)$, where $e_{n,t}$ is defined by

$$e_{n,t} = \begin{cases} \frac{Q_t - Q'_{n,t}}{|Q_t - Q'_{n,t}|} & \text{if } Q_t \neq Q'_{n,t}, \\ (1, 0, \dots, 0)^{\mathrm{T}} & \text{otherwise,} \end{cases}$$

for $Q_t \coloneqq X_t + V_t$ and $Q'_{n,t} \coloneqq X'_{n,t} + V'_{n,t}$, and B^{rc}_n, B^{sc}_n are defined by

$$\begin{aligned} \mathrm{d}B^{\mathrm{rc}}_{n,t} &= \mathrm{rc}_n(Z_t, Z'_{n,t})\,\mathrm{d}B_t + \mathrm{sc}_n(Z_t, Z'_{n,t})\,\mathrm{d}B''_t\,, \\ \mathrm{d}B^{\mathrm{sc}}_{n,t} &= \mathrm{sc}_n(Z_t, Z'_{n,t})\,\mathrm{d}B_t - \mathrm{rc}_n(Z_t, Z'_{n,t})\,\mathrm{d}B''_t\,. \end{aligned}$$

The solution Z'_n to this system of equations is well defined: they have Lipschitz-continuous coefficients in Z'_n, or in X'_n and V'_n (although the Lipschitz constants explode when $n \to +\infty$), so the existence and uniqueness of (strong) solution follow from Cauchy–Lipschitz arguments. By Lévy's characterization, B^{rc}_n, B^{sc}_n are independent Brownian motions and therefore if we define B'_n by

$$\mathrm{d}B'_{n,t} = \mathrm{rc}_n(Z_t, Z'_{n,t})(1 - 2e_{n,t}e_{n,t}^{\mathrm{T}})\,\mathrm{d}B^{\mathrm{rc}}_{n,t} + \mathrm{sc}_n(Z_t, Z'_{n,t})\,\mathrm{d}B^{\mathrm{sc}}_{n,t}\,,$$

then B'_n is also a Brownian motion. Hence Z'_n satisfies indeed (3.30) and it remains only to verify the last claim.

Difference process. Denote $\mathrm{rc}_{n,t} = \mathrm{rc}_n(Z_t, Z'_{n,t})$, $\mathrm{sc}_{n,t} = \mathrm{sc}_n(Z_t, Z'_{n,t})$, $\delta X_{n,t} = X_t - X'_{n,t}$, $\delta V_{n,t} = V_t - V'_{n,t}$, $\delta Q_{n,t} = Q_t - Q'_{n,t}$, $\delta g_{n,t} = g(Z_t) - g(Z'_{n,t})$, $\mathrm{d}W_{n,t} = e_{n,t}^{\mathrm{T}}\,\mathrm{d}B^{\mathrm{rc}}_{n,t}$ and $r_{n,t} = r(Z_t, Z'_{n,t}) = \theta|\delta X_{n,t}| + |\delta Q_{n,t}|$. In the following we will omit the subscript n in the variables defined above to simplify the notation and recover

[1]The additional Brownian motion B'' can be shared between stochastic processes Z'_n with different index n, so we do not need to extend the probability space infinitely.

it if necessary. By our construction of Brownian motions, the original Brownian B admits the decomposition

$$\mathrm{d}B_t = \mathrm{rc}(Z_t, Z_t')\,\mathrm{d}B_t^{\mathrm{rc}} + \mathrm{sc}(Z_t, Z_t')\,\mathrm{d}B_t^{\mathrm{sc}}\,.$$

Therefore, by taking the difference between the two systems of equations (3.29) and (3.30), we find the difference process $\delta Z = (\delta X, \delta V)$ satisfy

$$\begin{aligned}\mathrm{d}\delta X_t &= \delta V_t\,\mathrm{d}t\,,\\ \mathrm{d}\delta V_t &= (-\delta V_t - K\delta X_t + \delta g_t)\,\mathrm{d}t + 2\sqrt{2}\,\mathrm{rc}_t\,e_t\,\mathrm{d}W_t\,.\end{aligned}$$

We note that the process α disappears in the equations above. Using Itō's formula, we further obtain

$$\begin{aligned}\mathrm{d}(\delta X_t)^2 &= 2\delta X_t\cdot\delta V_t\,\mathrm{d}t\,,\\ \mathrm{d}(\delta V_t)^2 &= 2\delta V_t\cdot(-\delta V_t - K\delta X_t + \delta g_t)\,\mathrm{d}t + 8(\mathrm{rc}_t)^2\,\mathrm{d}t + 4\sqrt{2}\,\mathrm{rc}_t\,\delta V_t\cdot e_t\,\mathrm{d}W_t\,,\\ \mathrm{d}(\delta X_t\cdot\delta V_t) &= |\delta V_t|^2\,\mathrm{d}t + \delta X_t\cdot(-\delta V_t - K\delta X_t + \delta g_t)\,\mathrm{d}t + 2\sqrt{2}\,\mathrm{rc}_t\,\delta X_T\cdot e_t\,\mathrm{d}W_t\,.\end{aligned}$$

We have the semimartingle decomposition $\mathrm{d}r_t = \mathrm{d}A_t^r + \mathrm{d}M_t^r$ where A^r is absolutely continuous with

$$\mathrm{d}A_t^r \leqslant \big(|K| + L_1 - \theta\big)|\delta X_t|\,\mathrm{d}t + \theta|\delta Q_t|\,\mathrm{d}t\,,$$

and $M_t^r = 2\sqrt{2}\int_0^t \mathrm{rc}_s\,\mathrm{d}W_s$ is a martingale. Consequently, if $f : [0,+\infty) \to \mathbb{R}$ is a piecewise C^2, non-decreasing and concave function, then the Itō–Tanaka formula yields $\mathrm{d}f(r_t) = \mathrm{d}A_t^f + \mathrm{d}M_t^f$ where A^f is absolutely continuous with

$$\mathrm{d}A_t^f \leqslant f_-'(r_t)\big[\big(|K| + L_1 - \theta\big)|\delta X_t|\,\mathrm{d}t + \theta|\delta Q_t|\big]\,\mathrm{d}t + 4f''(r_t)(\mathrm{rc}_t)^2\,\mathrm{d}t\,,$$

and M^f is a martingale.

Choice of the Lyapunov function G. Define $G(z,z') = \frac{1}{2}\delta x^{\mathrm{T}}K\delta x + \frac{1}{2}|\delta v|^2 + \eta\delta x\cdot\delta v$, where we set $\eta = \frac{1}{2}\min(1,k)$, and denote $G_t = G(Z_t, Z_t')$. The function G satisfies the mutual bound

$$\lambda\big(|\delta x|^2 + |\delta v|^2\big) \leqslant G(z,z') \leqslant \frac{\theta}{2}\big(|\delta x|^2 + |\delta v|^2\big)\,,$$

where

$$\lambda \coloneqq \frac{1}{4}\min(1,k)\,.$$

We have also

$$\begin{aligned}\mathrm{d}G_t = &-\big((1-\eta)|\delta V_t|^2 + \eta\delta X_t\cdot\delta V_t + \eta\delta X_t^{\mathrm{T}}K\delta X_t\big)\,\mathrm{d}t\\ &+ (\delta V_t + \eta\delta X_t)\cdot(\delta g_t\,\mathrm{d}t + 2\sqrt{2}\,\mathrm{rc}_t\,e_t\,\mathrm{d}W_t) + 4(\mathrm{rc}_t)^2\,\mathrm{d}t \eqqcolon \mathrm{d}A_t^G + \mathrm{d}M_t^G\,,\end{aligned}$$

where A^G, M^G are the finite-variation and the martingale part respectively. In particular, if $|\delta X_t| + |\delta V_t| \geqslant R$, then

$$\begin{aligned}dA_t^G \leqslant -(\delta V_t,\ \delta X_t)\begin{pmatrix}1 - L_2 - \eta & -L_2 - (1+L_2)\eta\\ 0 & \eta(K - L_2)\end{pmatrix}\begin{pmatrix}\delta V_t\\ \delta X_t\end{pmatrix}\mathrm{d}t + 4(\mathrm{rc}_t)^2\,\mathrm{d}t\\ \eqqcolon -\delta Z_t^{\mathrm{T}}M_G\delta Z_t\,\mathrm{d}t + 4(\mathrm{rc}_t)^2\,\mathrm{d}t\,.\end{aligned}$$

We choose $\eta = \frac{1}{2}\min(1,k)$ and the symmetric part of the matrix M_G is positive-definite as its determinant is lower bounded by

$$(1-L_2-\eta)\cdot\eta(k-L_2) - \frac{1}{4}\big(L_2+(1+L_2)\eta\big)^2 \geqslant \frac{\big(\min(1,k)-19L_2\big)k}{16}\,.$$

By the same computation we obtain

$$M_G \succeq \frac{\big(\min(1,k)-19L_2\big)k}{16\max\big(1-L_2-\eta,\eta(k-L_2)\big)} \succeq \frac{\big(\min(1,k)-19L_2\big)k}{8\max\big(1-L_2,k-L_2\big)}\,.$$

As a result,

$$\begin{aligned}\mathrm{d}A^G_t - 4(\mathrm{rc}_t)^2\,\mathrm{d}t &\leqslant -\delta Z_t^{\mathrm{T}} M_G \delta Z_t\,\mathrm{d}t \leqslant -\frac{\big(\min(1,k)-19L_2\big)k}{8\max\big(1-L_2,k-L_2\big)}|\delta Z_t|^2\,\mathrm{d}t\\ &\leqslant -\frac{\big(\min(1,k)-19L_2\big)k}{8\max\big(1-L_2,k-L_2\big)\max\big(1,|K|\big)}G_t\,\mathrm{d}t =: -\kappa_2 G_t\,\mathrm{d}t\,.\end{aligned}$$

To summarize, if $|\delta X_t| + |\delta V_t| \geqslant R$, then $\mathrm{d}A^G_t \leqslant -\kappa_2 G_t\,\mathrm{d}t + 4(\mathrm{rc}_t)^2\,\mathrm{d}t$. In the general case, we have

$$\begin{aligned}\mathrm{d}A^G_t &\leqslant \big(|\delta V_t|,\ |\delta X_t|\big)\begin{pmatrix}|1-\eta|+L_1 & L_1+(1+L_1)\eta\\ 0 & \eta|K|+L_1\end{pmatrix}\begin{pmatrix}|\delta V_t|\\ |\delta X_t|\end{pmatrix}\mathrm{d}t + 4(\mathrm{rc}_t)^2\,\mathrm{d}t\\ &\leqslant \theta\big(|\delta X_t|^2+|\delta Y_t|^2\big)\mathrm{d}t + 4(\mathrm{rc}_t)^2\,\mathrm{d}t\,.\end{aligned}$$

Choice of ρ and f. Now recover the subscript n. Set

$$\rho_n(z,z') = \varepsilon_n G(z,z') + f_n\big(\theta|\delta x| + |\delta q|\big)$$

for $\varepsilon_n > 0$, and piecewise C^2, non-decreasing and concave $f_n : [0,+\infty) \to [0,+\infty)$, to be determined below. Denote $\rho_{n,t} = \rho_n(Z_{n,t}, Z'_{n,t})$. Then by the previous computations $\mathrm{d}\rho_{n,t} = \mathrm{d}A^{\rho_n}_t + \mathrm{d}M^{\rho_n}_t$ where M^{ρ_n} is a martingale and A^{ρ_n} is absolutely continuous with

$$\mathrm{d}A^{\rho_n}_t \leqslant \varepsilon_n\,\mathrm{d}A^G_t + \big[\big(|K|+L_1-\theta\big)|\delta X_{n,t}| + \theta|\delta Q_{n,t}|\big]f'_{n,-}(r_{n,t})\,\mathrm{d}t + 4f''_n(r_{n,t})(\mathrm{rc}_{n,t})^2\,\mathrm{d}t\,.$$

Define the functions

$$\begin{aligned}\varphi(r) &= \exp\bigg(-\frac{\theta r^2}{8}\bigg)\,,\\ \Phi(r) &= \int_0^r \varphi(u)\,\mathrm{d}u\,,\\ g_n(r) &= 1 - \frac{\kappa_{n,1}}{2}\int_0^r \Phi(u)\varphi(u)^{-1}\,\mathrm{d}u - \frac{\varepsilon_n}{2}\int_0^r \bigg[\bigg(1+\frac{\kappa_1}{2}\bigg)\theta u^2 + 4\bigg]\varphi(u)^{-1}\,\mathrm{d}u\end{aligned}$$

for $r \geqslant 0$, where $\kappa_{n,1}$, ε_n are positive constants defined by

$$\begin{aligned}\kappa_{n,1} &= \frac{1}{2}\bigg(\int_0^{r_0+n^{-1}} \Phi(u)\varphi(u)^{-1}\,\mathrm{d}u\bigg)^{-1}\,,\\ \varepsilon_n &= \frac{1}{2}\bigg(\int_0^{r_0+n^{-1}} \bigg[\bigg(1+\frac{\kappa_{n,1}}{2}\bigg)\theta u^2 + 4\bigg]\varphi(u)^{-1}\,\mathrm{d}u\bigg)^{-1} \wedge \frac{4}{9R}\,.\end{aligned}$$

We choose

$$f_n(r) = \int_0^{r\wedge(r_0+n^{-1})} \varphi(u)g(u)\,\mathrm{d}u\,.$$

Note that $\kappa_{n,1}$, ε_n, g_n and f_n all converge when $n \to +\infty$, and we denote their limit by

$$(\kappa_{*,1}, \varepsilon_*, g_*, f_*) = \lim_{n\to+\infty}(\kappa_1, \varepsilon, g, f)\,.$$

Denote also

$$\rho_*(z,z') = \varepsilon_* G(z,z') + f_*\big(\theta|\delta x| + |\delta q|\big)\,.$$

Since G is a quadratic form and $f_*'(0) = \varphi(0)g(0) = 1$, we have the uniform upper limit

$$\limsup_{z\to z'} \frac{\rho_*(z,z')}{|z-z'|} = \limsup_{z\to z'} \frac{\theta|x-x'| + |x-x'+v-v'|}{\sqrt{|x-x'|^2+|v-v'|^2}} \leqslant \theta + \sqrt{2} =: C_2\,,$$

validating the last property of the last claim. By elementary calculations, we also have

$$\lim_{n\to+\infty} \sup_{z,z'\in\mathbb{R}^{2d}} \frac{|\rho_n(z,z') - \rho_*(z,z')|}{1+|z-z'|^2} = 0\,. \tag{3.32}$$

The function f_n is C^2 on $[0, r_0+n^{-1})$ and $(r_0+n^{-1}, +\infty)$, non-decreasing, concave, and satisfies

$$4f_n''(r) + \theta f_n'(r)r + \kappa_{n,1}f_n(r) + \varepsilon_n\left[\left(1+\frac{\kappa_{n,1}}{2}\right)\theta r^2 + 4\right] \leqslant 0$$

for $r \in [0, r_0+n^{-1}]$. Moreover if $r \in [0, r_0+n^{-1}]$, then $\frac{1}{2} \leqslant g_n(r) \leqslant 1$. Therefore for every $r \geqslant 0$, we have

$$\frac{r}{2} \leqslant \frac{\Phi(r)}{2} \leqslant f_n(r) \leqslant \Phi(r)\,.$$

Proof of contraction. We now prove the contraction by investigating the following three cases. We temporarily omit the subscript n.

1. Suppose $r_t > r_0$. Then we have $|\delta X_t| + |\delta V_t| \geqslant R$ and therefore

$$\begin{aligned}
|\delta X_t|^2 + |\delta V_t|^2 \geqslant \frac{R^2}{2} &= \frac{R^2}{2f(r_0+n^{-1})} f(r_0+n^{-1}) \\
&\geqslant \frac{R^2}{2\Phi(r_0+n^{-1})} f(r_0+n^{-1}) \\
&\geqslant \frac{R^2}{2\Phi(r_0+n^{-1})} f(r_t)\,.
\end{aligned}$$

As a result, $G_t \geqslant \lambda R^2\big(2\Phi(r_0+n^{-1})\big)^{-1} f(r_t)$. Hence,

$$\begin{aligned}
\mathrm{d}A_t^\rho &\leqslant -\varepsilon\kappa_2 G_t\,\mathrm{d}t + \big(4f''(r_t) + \theta f_-'(r_t)r_t + 4\varepsilon\big)(\mathrm{rc}_t)^2\,\mathrm{d}t \leqslant -\varepsilon\kappa_2 G_t\,\mathrm{d}t \\
&\leqslant -\left(\varepsilon + \frac{2\Phi(r_0+n^{-1})}{\lambda R^2}\right)^{-1} \varepsilon\kappa_2\big(f(r_t) + \varepsilon G_t\big)\,\mathrm{d}t \\
&= -\frac{\lambda R^2\varepsilon}{\lambda R^2\varepsilon + 2\Phi(r_0+n^{-1})}\kappa_2\rho_t\,\mathrm{d}t\,.
\end{aligned}$$

2. Suppose $r_t \leqslant r_0$ and $|\delta Q_t| \geqslant 2n^{-1}$. Then we have $\mathrm{rc}_t = 1$ and therefore

$$\begin{aligned}\mathrm{d}A_t^\rho &\leqslant \varepsilon\theta\big(|\delta V_t|^2 + |\delta X_t|^2\big)\,\mathrm{d}t + \big(4f''(r_t) + \theta f'_-(r_t)r_t + 4\varepsilon\big)\,\mathrm{d}t \\ &\leqslant \varepsilon\theta r_t^2\,\mathrm{d}t + \big(4f''(r_t) + \theta f'_-(r_t)r_t + 4\varepsilon\big)\,\mathrm{d}t \\ &\leqslant \big(4f''(r_t) + \theta f'_-(r_t)r_t + \varepsilon(\theta r_t^2 + 4)\big)\,\mathrm{d}t \\ &\leqslant -\kappa_1\left(f(r_t) + \frac{\varepsilon\theta}{2}r_t^2\right)\mathrm{d}t \\ &\leqslant -\kappa_1\big(f(r_t) + \varepsilon G_t\big)\,\mathrm{d}t = -\kappa_1\rho_t\,\mathrm{d}t\,.\end{aligned}$$

3. Suppose $r_t \leqslant r_0$ and $|\delta Q_t| < 2n^{-1}$. Then we have

$$|\delta V_t|^2 \leqslant 2|\delta X_t|^2 + 2|\delta Q_t|^2 \leqslant 2|\delta X_t|^2 + 8n^{-2}\,.$$

and

$$|\delta X_t| \leqslant \theta^{-1}r_t \leqslant \theta^{-1}r_0 = \frac{\theta+1}{\theta}R \leqslant \frac{3}{2}R\,.$$

Consequently,

$$\begin{aligned}\mathrm{d}A_t^\rho &\leqslant \varepsilon\theta\big(|\delta V_t|^2 + |\delta X_t|^2\big)\,\mathrm{d}t - \frac{\theta}{2}|\delta X_t|\,\mathrm{d}t + 2\theta n^{-1}\,\mathrm{d}t + \big(4f''(r_t) + 4\varepsilon\big)(\mathrm{rc}_t)^2\,\mathrm{d}t \\ &\leqslant \varepsilon\theta\big(3|\delta X_t|^2 + 8n^{-2}\big)\,\mathrm{d}t - \frac{\theta}{2}|\delta X_t|\,\mathrm{d}t + 2\theta n^{-1}\,\mathrm{d}t \\ &\leqslant -\frac{\theta}{4}|\delta X_t|\,\mathrm{d}t + (2\theta n^{-2} + 8\theta\varepsilon n^{-2})\,\mathrm{d}t\,.\end{aligned}$$

Since

$$f(r_t) \leqslant \sup_{r\in(0,r_0]}\frac{f(r)}{r}r_t = r_t \leqslant \theta|\delta X_t| + 2n^{-1}\,,$$

$$G_t \leqslant \theta\big(|\delta X_t|^2 + |\delta V_t|^2\big) \leqslant \theta|\delta X_t|^2 + 4\theta n^{-2} \leqslant \frac{3\theta R}{2}|\delta X_t| + 4\theta n^{-2}\,,$$

we have

$$\begin{aligned}\mathrm{d}A_t^\rho &\leqslant -\frac{1}{4+6\varepsilon R}\big(f(r_t) + \varepsilon G_t - 2n^{-1} - 4\theta\varepsilon n^{-1}\big)\,\mathrm{d}t + (2\theta n^{-2} + 8\theta\varepsilon n^{-2})\,\mathrm{d}t \\ &= -\frac{\rho_t}{4+6\varepsilon R}\,\mathrm{d}t + O(n^{-1})\,\mathrm{d}t\,.\end{aligned}$$

Recovering the subscript n and combining the three cases above, we obtain

$$\mathbb{E}\big[\rho_n(Z_t, Z'_{n,t})\big] = \mathbb{E}[\rho_{n,t}] \leqslant e^{-\kappa_n t}\,\mathbb{E}[\rho_{n,0}] + O(n^{-1}) = e^{-\kappa_n t}\,\mathbb{E}\big[\rho_n(Z_0, Z'_0)\big] + O(n^{-1})$$

for

$$\kappa_n := \min\left(\kappa_{n,1}, \frac{\lambda R^2\varepsilon_n\kappa_2}{\lambda R^2\varepsilon_n + 2\Phi(r_0 + n^{-1})}, \frac{1}{4+6\varepsilon_n R}\right).$$

Thanks to the uniform convergence of ρ_n to ρ_* in (3.32) and the square-integrability of the processes Z_t, $Z'_{n,t}$, we can take the limit $n \to +\infty$ to derive

$$\limsup_{n\to+\infty}\mathbb{E}\big[\rho_*(Z_t, Z'_{n,t})\big] = \limsup_{n\to+\infty}\mathbb{E}[\rho_{*,t}] \leqslant e^{-\kappa_* t}\,\mathbb{E}[\rho_{*,0}] = e^{-\kappa_* t}\,\mathbb{E}\big[\rho_*(Z_0, Z'_0)\big]$$

for

$$\begin{aligned}\kappa &\coloneqq \min\left(\kappa_{*,1}, \frac{\lambda R^2 \varepsilon_* \kappa_2}{\lambda R^2 \varepsilon_* + 2\Phi\big((\theta+1)R\big)}, \frac{1}{4+6\varepsilon_* R}\right) \\ &\geqslant \min\left(\kappa_{*,1}, \frac{\lambda R^2 \varepsilon_* \kappa_2}{\lambda R^2 \varepsilon_* + 2\Phi\big((\theta+1)R\big)}, \frac{3}{20}\right),\end{aligned}$$

where the inequality is due to $\varepsilon_* \leqslant \frac{4}{9R}$. If $r(z,z') \leqslant r_0$, then

$$\begin{aligned}|z-z'| &\leqslant \sqrt{2}\big(|x-x'| + |v-v'|\big) \\ &\leqslant \sqrt{2}\big(2|x-x'| + |q-q'|\big) \\ &\leqslant \sqrt{2}\, r \leqslant \frac{2\sqrt{2}\, r_0}{\Phi(r_0)} f_*(r)\,;\end{aligned}$$

otherwise $|\delta x| + |\delta v| \geqslant R$, and then

$$|z-z'| = \frac{|\delta x|^2 + |\delta v|^2}{|z-z'|} \leqslant \frac{\sqrt{2}}{\lambda R} G(z,z')\,.$$

Therefore for every z, $z' \in \mathbb{R}^{2d}$, we have

$$|z-z'| \leqslant C_1\big[\varepsilon_* G(z,z') + f_*\big(\theta|x-x'| + |q-q'|\big)\big] = C_1 \rho_*(z,z')$$

with

$$C_1 = \sqrt{2}\max\left(\frac{2(\theta+1)R}{\Phi\big((\theta+1)R\big)}, \frac{1}{\lambda\varepsilon_* R}\right).$$

As a consequence,

$$\limsup_{n\to+\infty} \mathbb{E}\big[|Z_t - Z'_{n,t}|\big] \leqslant \limsup_{n\to+\infty} \mathbb{E}\big[\rho_*(Z_t, Z'_{n,t})\big] \leqslant C_1 e^{-\kappa t}\, \mathbb{E}\big[\rho(Z_0, Z'_0)\big]\,. \qquad \square$$

Chapter 4

Time-uniform log-Sobolev inequalities and applications to propagation of chaos

Abstract. Time-uniform log-Sobolev inequalities (LSI) satisfied by solutions of semi-linear mean-field equations have recently appeared to be a key tool to obtain time-uniform propagation of chaos estimates. This work addresses the more general settings of time-inhomogeneous Fokker–Planck equations. Time-uniform LSI are obtained in two cases, either with the bounded-Lipschitz perturbation argument with respect to a reference measure, or with a coupling approach at high temperature. These arguments are then applied to mean-field equations, where, on the one hand, sharp marginal propagation of chaos estimates are obtained in smooth cases and, on the other hand, time-uniform global propagation of chaos is shown in the case of vortex interactions with quadratic confinement potential on the whole space. In this second case, an important point is to establish global gradient and Hessian estimates, which is of independent interest. We prove these bounds in the more general situation of non-attractive logarithmic and Riesz singular interactions.

Based on joint work with Pierre Monmarché and Zhenjie Ren.

4.1 Introduction

We are interested in families $(m_t)_{t\geqslant 0}$ of probability distributions solving time-inhomogeneous Fokker–Planck equations on $\mathbb{R}^d$ of the form

$$\partial_t m_t = \nabla \cdot (\sigma^2 \nabla m_t - b_t m_t)\,, \tag{4.1}$$

where $\sigma^2 > 0$ and $b_t : \mathbb{R}^d \to \mathbb{R}^d$ for $t \geqslant 0$. This describes the evolution of the law of the diffusion process

$$\mathrm{d}X_t = b_t(X_t)\,\mathrm{d}t + \sqrt{2}\sigma\,\mathrm{d}B_t\,, \tag{4.2}$$

where B is a standard d-dimensional Brownian motion. We have particularly in mind McKean–Vlasov equations, where b_t is in fact a function of m_t itself, namely

$$b_t(x) = F(x, m_t)\,, \tag{4.3}$$

for some suitable function F. Other examples are time-integrated McKean–Vlasov equations where $b_t(x) = F\big(x, \int_0^t m_s k_t(\mathrm{d}s)\big)$ for some kernel k_t (as in [46]).

Denoting by $\mathcal{C}^1_{\mathrm{c}}(\mathbb{R}^d)$ the set of compactly supported $\mathcal{C}^1$ functions from $\mathbb{R}^d$ to $\mathbb{R}$, a probability measure μ on $\mathbb{R}^d$ is said to satisfy a log-Sobolev inequality (LSI) with constant $C > 0$ if

$$\forall h \in \mathcal{C}^1_{\mathrm{c}}(\mathbb{R}^d) \text{ with } \int_{\mathbb{R}^d} h^2 \,\mathrm{d}\mu = 1\,, \qquad \int_{\mathbb{R}^d} h^2 \ln(h^2)\,\mathrm{d}\mu \leqslant C \int_{\mathbb{R}^d} |\nabla h|^2 \,\mathrm{d}\mu\,. \quad (4.4)$$

Equivalently, for all probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$ such that ν is absolutely continuous with respect to μ and $\sqrt{\mathrm{d}\nu/\mathrm{d}\mu} \in \mathcal{C}^1_{\mathrm{c}}$, we have

$$\mathcal{H}(\nu|\mu) \leqslant \frac{C}{4}\mathcal{I}(\nu|\mu)\,,$$

where $\mathcal{H}$, $\mathcal{I}$ are the relative entropy and Fisher information defined respectively as follows:

$$\mathcal{H}(\nu|\mu) \coloneqq \int_{\mathbb{R}^d} \ln\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\nu\,,$$
$$\mathcal{I}(\nu|\mu) \coloneqq \int_{\mathbb{R}^d} \left|\nabla \ln\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right|^2 \mathrm{d}\nu\,.$$

We want to determine suitable conditions under which the family $(m_t)_{t\geqslant 0}$ solving (4.1) satisfies a uniform LSI, in the sense that (4.4) holds with $\mu = m_t$ and a constant C independent from t. As will be discussed below in details (in Sections 4.3 and 4.4), for McKean–Vlasov equations, this is an important tool to get uniform-in-time Propagation of Chaos (PoC) estimates [98, 142].

The paper is organized as follows. In the rest of this introduction we state our main results concerning time-uniform LSI (Theorem 4.3 and 4.4), which are proven in Section 4.2. In Section 4.3 we use them to extend the range of the work [142] of Lacker and Le Flem, obtaining sharp uniform in time PoC for McKean–Vlasov equations in cases of smooth interaction. Section 4.4 addresses the question of uniform-in-time LSI and PoC for singular (log or Riesz) interactions in $\mathbb{R}^d$.

Before stating our main results, we recall first the following result of Malrieu [159], based on the classical Bakry–Émery approach.

Proposition 4.1. *Assume that there exist $T > 0$, $L \in \mathbb{R}$ such that for all $t \in [0, T]$ and x, $y \in \mathbb{R}^d$,*

$$\big(b_t(x) - b_t(y)\big) \cdot (x - y) \leqslant L|x - y|^2\,, \quad (4.5)$$

and that m_0 satisfies an LSI with constant $C_0 > 0$. Then, for all $t \in [0, T]$, the measure m_t satisfies an LSI with constant

$$C_t = e^{2Lt} C_0 + \sigma^2 \int_0^t e^{2Ls}\,\mathrm{d}s\,.$$

For completeness, the proof is recalled in Section 4.2.1.

Remark 4.2. When the curvature lower-bound L in (4.5) is negative, this already gives an LSI uniform in t, but we are mostly interested in cases where (4.5) only holds with $L > 0$. Nevertheless, this first proposition means that, in the next results (Theorems 4.3 and 4.4), in fact, if the assumptions are only satisfied for $t \geqslant t_0$ for some $t_0 > 0$ large enough (for instance the condition (4.8)), we can apply Proposition 4.1 for times $t \in [0, t_0]$ and then apply the other results to $(m_{t+t_0})_{t\geqslant 0}$.

The next result addresses the high-diffusivity regime, namely when σ^2 is high enough (see (4.8)). It is proven in Section 4.2.2.

Theorem 4.3. *Assume that there exist ρ, L, R, $K > 0$ such that, for all $t \geqslant 0$,*

$$\big(b_t(x) - b_t(y)\big) \cdot (x-y) \leqslant \begin{cases} -\rho|x-y|^2 & \forall x, y \in \mathbb{R}^d \text{ with } |x| \geqslant R\,, \\ L|x-y|^2 & \forall x, y \in \mathbb{R}^d\,, \end{cases} \tag{4.6}$$

and, setting $R_* = R(2 + 2L/\rho)^{1/d}$,

$$\sup_{|x| \leqslant R_*} \{-x \cdot b_t(x)\} \leqslant K\,. \tag{4.7}$$

Then, provided m_0 satisfies an LSI and

$$\sigma^2 \geqslant \sigma_0^2 \coloneqq 2(2L+\rho)\frac{(L+\rho/4)R_*^2 + K}{\rho d}\,, \tag{4.8}$$

the family $(m_t)_{t\geqslant 0}$ satisfies a uniform LSI.

Moreover, there exists $C_* > 0$ which depends on L, R, d and ρ but not on m_0, K nor σ such that, provided (4.8), for t large enough, the measure m_t satisfies an LSI with constant $\sigma^2 C_*$.

More precisely, for any $\varepsilon > 0$, there exists $\sigma_0' > 0$ which depends only on L, R, d, ρ and ε such that for all $\sigma \geqslant \sigma_0'$, for t large enough, the measure m_t satisfies an LSI with constant $\sigma^2(\rho^{-1} + \varepsilon)$.

The next result is the adaptation in the time-inhomogeneous settings of the bounded-Lipschitz perturbation argument of Chapter 3. Its proof is given in Section 4.2.3.

Theorem 4.4. *Assume that, for all $t \geqslant 0$, the drift writes $b_t = a_0 + g_t$ for some a_0, $g_t \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$ with bounded derivatives such that the generator $\mathcal{L}_0 = a_0 \cdot \nabla + \sigma^2 \Delta$ admits a unique C^2 invariant probability density μ_0 satisfying an LSI. Write $\tilde{b}_t \coloneqq 2\nabla \ln \mu_0 - b_t$ and $\varphi_t \coloneqq -\nabla \cdot g_t + g_t \cdot \nabla \ln \mu_0$. Assume that there exist L, R, M^φ, $L^\varphi \geqslant 0$ and $\rho > 0$ such that, for all $t \geqslant 0$, we have $\varphi_t = \varphi_{1,t} + \varphi_{2,t}$ with M^φ-bounded $\varphi_{1,t}$ and L^φ-Lipschitz $\varphi_{2,t}$ and for all x, $y \in \mathbb{R}^d$,*

$$\big(\tilde{b}_t(x) - \tilde{b}_t(y)\big) \cdot (x-y) \leqslant \begin{cases} -\rho|x-y|^2 & \text{if } |x-y| \geqslant R, \\ L|x-y|^2 & \text{otherwise.} \end{cases} \tag{4.9}$$

Finally, assume that m_0 admits a density e^{u_0} with respect to μ_0, with u_0 being the sum of a bounded and a Lipschitz continuous functions. Then $(m_t)_{t\geqslant 0}$ satisfies a uniform LSI.

Moreover there exists $C_* > 0$ which depends on L, R, M^φ, L^φ, σ^2, ρ and the LSI constant of μ_0 but not on m_0 such that, for some $t_* > 0$, m_t satisfies an LSI with constant C_* for all $t \geqslant t_*$.

Finally, denoting by C_0 the LSI constant of μ_0, the following holds. For any $\varepsilon > 0$, there exists $\eta > 0$ (which depends only on ρ, L, R and ε) such that, if $M^\varphi + L^\varphi \leqslant \eta$, there exists t_* such that m_t satisfies an LSI with constant $C_0 + \varepsilon$ for all $t \geqslant t_*$.

4.2 Proofs of the general results

In this section we write $(P_{s,t})_{t\geqslant s\geqslant 0}$ the inhomogeneous Markov semi-group associated to (4.2), given by

$$P_{s,t}f(x) = \mathbb{E}[f(X_t)|X_s = x] .$$

In particular, the solution of (4.1) is then given by $m_t = m_0 P_{0,t}$. In the proofs of Proposition 4.1 and Theorem 4.3, we can additionally assume that b_t is smooth with all derivatives being bounded, and consider functions f which are for instance the sum of a positive constant and a compactly supported smooth non-negative function, which enable to justify the computations based on $\partial_t P_{s,t} f = P_{s,t}\mathcal{L}_t f$ and $\partial_s P_{s,t} f = -\mathcal{L}_s P_{s,t} f$ (using e.g. Proposition C.2). The conclusion is then obtained by approximation (as in e.g. Chapter 1).

4.2.1 Proof of Proposition 4.1

Proof of Proposition 4.1. Considering X and X' two solutions of (4.2) driven by the same Brownian motion, the condition (4.5) gives

$$\mathrm{d}|X_t - X'_t|^2 \leqslant 2L|X_t - X'_t|^2 \,\mathrm{d}t , \tag{4.10}$$

so that $|X_t - X'_t|^2 \leqslant e^{2L(t-s)}|X_s - X'_s|^2$ for all $t \geqslant s \geqslant 0$, which by [139] implies

$$|\nabla P_{s,t} f| \leqslant e^{L(t-s)} P_{s,t}|\nabla f| . \tag{4.11}$$

Fix a function $f \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R}_+)$, globally Lipschitz continuous and lower bounded by a positive constant (it is sufficient to prove the LSI with these functions and conclude by approximation). For $t > s \geqslant 0$, we consider the interpolation $\Psi(u) = P_{s,u}(P_{u,t} f \ln P_{u,t} f)$ for $u \in [s, t]$, so that

$$\begin{aligned} P_{s,t}(f \ln f) - P_{s,t} f \ln P_{s,t} f &= \Psi(t) - \Psi(s) \\ &= \int_s^t \Psi'(u) \,\mathrm{d}u \\ &= \sigma^2 \int_s^t P_{s,u} \frac{|\nabla P_{u,t} f|^2}{P_{u,t} f} \,\mathrm{d}u \\ &\leqslant \sigma^2 \int_s^t e^{2L(t-u)} \,\mathrm{d}u P_{s,t}\left(\frac{|\nabla f|^2}{f}\right) , \end{aligned}$$

where we used that $(P_{u,t}|\nabla f|)^2 \leqslant P_{u,t}(|\nabla f|^2/f) P_{u,t}(f)$ by Cauchy–Schwarz. Integrating with respect to m_s gives

$$m_t(f \ln f) \leqslant m_s(P_{s,t} f \ln P_{s,t} f) + m_t\left(\frac{|\nabla f|^2}{f}\right)\sigma^2 \int_s^t e^{2L(t-u)} \,\mathrm{d}u . \tag{4.12}$$

The proof is concluded by applying this with $s = 0$ and using the LSI for m_0, (4.11) and Cauchy–Schwarz to bound

$$\begin{aligned} m_0(P_{0,t} f \ln P_{0,t} f) &\leqslant m_t f \ln(m_t f) + C_0 m_0\left(\frac{|\nabla P_{0,t} f|^2}{P_{0,t} f}\right) \\ &\leqslant m_t f m_t(\ln f) + C_0 e^{2Lt} m_t\left(\frac{|\nabla f|^2}{f}\right) . \end{aligned}$$

□

4.2.2 Proof of Theorem 4.3

Proof of Theorem 4.3. The different steps of the proof are the following. First, using the coupling argument of [168] (at high diffusivity), we get a long-time L^2 contraction along the synchronous coupling of two solutions of (4.2). By contrast to the almost sure contraction (4.10), this L^2 contraction is not enough to get an LSI, but it gives a uniform Poincaré inequality following arguments similar to the proof of Proposition 4.1. It remains then to prove a so-called defective LSI, which together with the Poincaré inequality yields the desired LSI. The proof of the defective LSI follows the arguments of Chapter 3, except that in the present case the measure for which the LSI is proven is not an invariant measure of a time-homogeneous semi-group (which would solve $\mu = \mu P_t$, which in our case is replaced by $m_t = m_0 P_{0,t}$). These arguments combine a Wang–Harnak inequality for the operator $P_{0,t}$ with a Gaussian moment bound.

Step 1: Poincaré inequality. Let X, X' be two solutions of (4.2) driven by the same Brownian motion. Following the proof of [168, Theorem 1] (which is concerned with time-homogeneous processes, but the proof works readily in the non-homogeneous case under the time-uniform assumptions made in Theorem 4.3), we get for all $t \geqslant s \geqslant 0$,

$$\mathbb{E}\big[|X_t - X_t'|^2\big] \leqslant M e^{-\lambda(t-s)} \mathbb{E}\big[|X_s - X_s'|^2\big] \,,$$

where

$$\lambda = \frac{\rho}{2} \,, \qquad M = 1 + \frac{2(2L+\rho)R_*^2}{4d\sigma^2} \,. \tag{4.13}$$

This implies, by [139],

$$|\nabla P_{s,t} f|^2 \leqslant M e^{-\lambda(t-s)} P_{s,t} |\nabla f|^2 \,.$$

Since m_0 satisfies a LSI, it satisfies a Poincaré inequality, and thus,

$$m_0 (P_{0,t} f)^2 - (m_0 P_{0,t} f)^2 \leqslant C_0 m_0 |\nabla P_{0,t} f|^2 \leqslant C_0 M e^{-\lambda t} m_t |\nabla f|^2 \,.$$

Besides, for fixed $t \geqslant 0$ and $f \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ globally Lipschitz continuous, considering the interpolation $\Psi(u) = P_{0,u}(P_{u,t} f)^2$ for $u \in [0,t]$, we get

$$\begin{aligned} P_{0,t}(f^2) - (P_{0,t} f)^2 &= \Psi(t) - \Psi(0) \\ &= \int_0^t \Psi'(u)\,\mathrm{d}u \\ &= \sigma^2 \int_0^t P_{0,u} |\nabla P_{u,t} f|^2 \,\mathrm{d}u \\ &\leqslant \sigma^2 \int_0^t M e^{-\lambda(t-u)} \,\mathrm{d}u P_{0,t} |\nabla f|^2 \,. \end{aligned}$$

Combining these last two inequalities, we get

$$\begin{aligned} m_t(f^2) - \big(m_t(f)\big)^2 &= m_0\big(P_{0,t}(f^2) - (P_{0,t} f)^2\big) + m_0 (P_{0,t} f)^2 - (m_0 P_{0,t} f)^2 \\ &\leqslant M \left(\frac{\sigma^2}{\lambda} + e^{-\lambda t} C_0 \right) m_t |\nabla f|^2 \,, \end{aligned} \tag{4.14}$$

which is a uniform Poincaré inequality for $(m_t)_{t \geqslant 0}$.

Step 2: Gaussian moment. Since m_0 satisfies an LSI, there exists $\delta_0 > 0$ such that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{\delta_0|x-y|^2} m_0(\mathrm{d}x)m_0(\mathrm{d}y) < \infty\,.$$

Write $V(x,y) = e^{\delta|x-y|^2}$ for some $0<\delta<\min(\delta_0,\rho/5)$ and $\mathcal{L}_{2,t}$ the generator on $\mathbb{R}^d\times\mathbb{R}^d$ of two independent diffusion processes (4.2), namely

$$\mathcal{L}_{2,t}g(x,y) = b_t(x)\cdot\nabla_x + b_t(y)\cdot\nabla_y + \sigma^2\Delta_x + \sigma^2\Delta_y\,.$$

Using (4.6) (and that $|x-y|\geqslant 2R$ implies that either $|x|\geqslant R$ or $|y|\geqslant R$),

$$\begin{aligned}\frac{\mathcal{L}_{2,t}V(x,y)}{V(x,y)} &= 2\delta(x-y)\cdot\big(b_t(x)-b_t(y)\big) + 4\delta d + 8\delta^2|x-y|^2\\ &\leqslant \begin{cases}4\delta d + (8\delta^2-2\delta\rho)|x-y|^2 & \text{if } |x-y|\geqslant 2R\\ 4\delta d + 4(8\delta^2+2\delta L)R^2 & \text{otherwise}\end{cases}\\ &\leqslant -\delta\mathbb{1}_{|x-y|\geqslant R_*} + C_*\mathbb{1}_{|x-y|<R_*}\end{aligned}$$

with

$$R_*^2 = \max\left(\frac{1+4d}{2(\rho-4\delta)}, 4R^2\right), \qquad C_* = \delta\big(4d + (2L+8\delta)R^2\big)\,.$$

Hence,

$$\mathcal{L}_{2,t}V(x,y) \leqslant -\delta V(x,y) + C_* e^{\delta R_*^2}\,,$$

and thus,

$$\partial_t(m_t\otimes m_t)(V) \leqslant -\delta(m_t\otimes m_t)(V) + C_* e^{\delta R_*^2}\,.$$

As a conclusion, for all $t\geqslant 0$,

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{\delta|x-y|^2} m_t(\mathrm{d}x)m_t(\mathrm{d}y) \leqslant \delta^{-1}C_* e^{\delta R_*^2} + e^{-\delta t}\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{\delta|x-y|^2} m_0(\mathrm{d}x)m_0(\mathrm{d}y)\,. \tag{4.15}$$

Step 3: Wang–Harnack inequality. In the following, fix $f>0$ such that $m_t f = 1$. Using the Röckner–Wang argument for the diffusion (4.2) we get, for all x, $y\in\mathbb{R}^d$, $\alpha>1$ and $t>0$,

$$\big(P_{0,t}f(y)\big)^\alpha \leqslant (P_{0,t}f^\alpha)(x)\exp\left(\frac{\alpha}{2\sigma^2(\alpha-1)}\left(L^2t + \frac{|x-y|^2}{t}\right)\right), \tag{4.16}$$

so that

$$\begin{aligned}&\int_{\mathbb{R}^d}(P_{0,t}f^\alpha)(x)m_0(\mathrm{d}x)\\ &\qquad\geqslant \big(P_{0,t}f(y)\big)^\alpha\int_{\mathbb{R}^d}\exp\left(-\frac{\alpha}{2\sigma^2(\alpha-1)}\left(L^2t+\frac{|x-y|^2}{t}\right)\right)m_0(\mathrm{d}x)\,,\end{aligned}$$

and thus, for any $\beta>\alpha$,

$$\begin{aligned}&m_0(P_{0,t}f)^\beta\\ &\quad\leqslant \big(m_t(f^\alpha)\big)^{\beta/\alpha}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\exp\left(-\frac{\alpha\big(L^2t+\frac{|x-y|^2}{t}\big)}{2\sigma^2(\alpha-1)}\right)m_0(\mathrm{d}x)\right]^{-\beta/\alpha}m_0(\mathrm{d}y)\\ &\quad\leqslant \big(m_t(f^\alpha)\big)^{\beta/\alpha}\int_{\mathbb{R}^{2d}}\exp\left(\frac{\beta\big(L^2t+\frac{|x-y|^2}{t}\big)}{2\sigma^2(\alpha-1)}\right)m_0(\mathrm{d}x)m_0(\mathrm{d}y)\,.\end{aligned}\tag{4.17}$$

Using Jensen's inequality for the probability with density $P_{0,t}f$ with respect to m_0, taking $\alpha = 3/2$ so that $x \mapsto x^{\alpha-1}$ is concave, we get

$$m_t(f^\alpha) = m_0 P_{0,t} f^\alpha \leqslant \big(m_0(P_{0,t}f)^2\big)^{\alpha-1}.$$

Using (4.17) with $\beta = 2$ to bound the right hand side then gives

$$m_t(f^\alpha) \leqslant \big(m_t(f^\alpha)\big)^{2/3}\left[\int_{\mathbb{R}^{2d}} \exp\left(\frac{4}{2\sigma^2}\left(L^2 t + \frac{|x-y|^2}{t}\right)\right) m_0(\mathrm{d}x) m_0(\mathrm{d}y)\right]^{1/2}.$$

and we can divide by $(m_t f^\alpha)^{2/3}$ to end up with

$$m_t(f^\alpha) \leqslant \left[\int_{\mathbb{R}^{2d}} \exp\left(\frac{2}{\sigma^2}\left(L^2 t + \frac{|x-y|^2}{t}\right)\right) m_0(\mathrm{d}x) m_0(\mathrm{d}y)\right]^{3/2}.$$

Applying this result to $(m_{t+t_0})_{t\geqslant 0}$ for some $t_0 > 0$ we get that for all $t \geqslant 0$ and all $f > 0$ with $m_{t+t_0} f = 1$,

$$m_{t+t_0}\big(f^{3/2}\big) \leqslant \left[\int_{\mathbb{R}^{2d}} \exp\left(\frac{2}{\sigma^2}\left(L^2 t_0 + \frac{|x-y|^2}{t_0}\right)\right) m_t(\mathrm{d}x) m_t(\mathrm{d}y)\right]^{3/2}.$$

Taking $t_0 = 2/(\delta\sigma^2)$, the right hand side is bounded uniformly in $t \geqslant 0$ thanks to (4.15). As a conclusion, we have determined t_0, $C > 0$ such that

$$\forall t \geqslant t_0,\ \forall f > 0, \qquad m_t\big(f^{3/2}\big) \leqslant C(m_t f)^{3/2}. \tag{4.18}$$

Moreover, in view of (4.15), we can find $C > 0$ which depends on m_0 only through δ such that (4.18) holds with this C for all t large enough. To see that we can take δ independent from m_0, we can replace the function V above by the time-dependent $V_t(x,y) = e^{\delta_t|x-y|^2}$ where $t \mapsto \delta_t$ is slowly and smoothly increasing starting from some small $\delta_0 > 0$ (depending on m_0) and reaching $\rho/5$ after some time. Following similar computations as above we get that $(m_t \otimes m_t)(V_t)$ is non-increasing (taking $\mathrm{d}\delta_t/\mathrm{d}t$ sufficiently small), from which, replacing $(m_t)_{t\geqslant 0}$ by $(m_{t_0+t})_{t\geqslant 0}$ for some sufficiently large t_0, we can assume that (4.15) holds for $\delta = \rho/5$. As a conclusion, for times large enough, (4.18) holds with a constant C independent from m_0.

Step 4: Conclusion. For $t \geqslant t_0$, applying (4.12) with $s = t - t_0$ gives, for $f > 0$,

$$m_t(f \ln f) \leqslant 2 m_s\big(P_{s,t} f \ln(P_{s,t} f)^{1/2}\big) + m_t\left(\frac{|\nabla f|^2}{f}\right)\sigma^2 \int_0^{t_0} e^{2Lu}\,\mathrm{d}u.$$

Assume that f is such that $m_t f = 1$. Applying Jensen's inequality twice (first with the probability measure with density $P_{s,t} f$ with respect to m_s) gives

$$m_s\big(P_{s,t} f \ln(P_{s,t} f)^{1/2}\big) \leqslant \ln\big(m_s(P_{s,t} f)^{3/2}\big) \leqslant \ln m_t\big(f^{3/2}\big).$$

Thanks to (4.18), we have thus obtained that for all $t \geqslant t_0$ and all $f > 0$ with $m_t f = 1$,

$$m_t(f \ln f) \leqslant 2 \ln C + m_t\left(\frac{|\nabla f|^2}{f}\right)\sigma^2 \int_0^{t_0} e^{2Lu}\,\mathrm{d}u, \tag{4.19}$$

which is called a defective LSI (and is uniform over $t \geqslant t_0$). According to [12, Proposition 5.1.3], combining this inequality with the (time uniform) Poincaré inequality (4.14) gives an LSI for m_t uniformly over $t \geqslant t_0$. For $t \in [0, t_0]$ we apply Proposition 4.1, which concludes the proof of the uniform LSI.

Finally, as mentioned above, the constant C may be taken independent from m_0, in which case the defective LSI (4.19) holds for sufficiently large times. Similarly, we see that the Poincaré inequality (4.14) holds with constant $M\sigma^2/\lambda + 1$ (which is independent from m_0) for t large enough. This shows that there exists $C'_* > 0$ independent from m_0 such that m_t satisfies an LSI with constant C'_* for t large enough. The fact that $C'_* \leqslant \sigma^2 C_*$ for some $C_* > 0$ independent from σ can be checked in the explicit expressions above. More precisely, taking $\delta = \rho/5$ and $t_0 = 2/(\delta\sigma^2)$, we get that, in (4.19) the constant C is uniformly bounded over $\sigma \geqslant \sigma_0$ by a constant that depends only on ρ, L, R, d, and similarly we can bound

$$\sigma^2 \int_0^{t_0} e^{2Lu}\,\mathrm{d}u \leqslant \sigma^2 t_0 e^{2Lt_0} \leqslant \frac{10}{\rho} e^{20L/\rho}$$

in (4.19) uniformly over $\sigma \geqslant 1$. As a consequence, for large values of σ^2, the leading term in the LSI constant for large times is $\sigma^2 M/\lambda$ from the Poincaré constant, with M and λ in (4.13). As $\sigma \to \infty$, M goes to 1, so we may take the LSI constant (for large times) to be $\sigma^2(\lambda^{-1} + \varepsilon)$ for any arbitrary $\varepsilon > 0$ for σ large enough. This estimate (with $\lambda = \rho/2$) is not sharp, as we expect an LSI of order σ^2/ρ (which is the Gaussian behavior). This is due to the $1/2$ factor in the definition of λ in [168], which is in fact arbitrary, in the sense that the computations of [168] work if we take $\lambda = \alpha\rho$ for an arbitrary $\alpha < 1$ (see the two first equations of [168, Section 2.1.2]), provided the lower bound on the temperature σ_0^2 is sufficiently large (depending on α). As a conclusion, we can get a Poincaré constant, and thus an LSI constant, equal to $\sigma^2(\rho^{-1} + \varepsilon)$ for an arbitrary ε for large times, provided σ is large enough. $\square$

4.2.3 Proof of Theorem 4.4

Proof of Theorem 4.4. The proof closely follows the one of Theorem 3.1 (in the time-homogeneous settings and with $m_0 = \mu_0$, i.e. $u_0 = 0$), the time dependencies appearing along the proof being dealt with the uniform-in-time assumptions of Theorem 4.4. We recall the main arguments and refer to Chapter 3 for details. Starting from

$$\begin{aligned} \partial_t m_t &= -\nabla \cdot (b_t m_t) + \Delta m_t\,, \\ 0 = \partial_t \mu_0 &= -\nabla \cdot (a_0 \mu_0) + \Delta \mu_0\,, \end{aligned}$$

we get that $h_t = m_t/\mu_0$ is a viscosity solution to

$$\partial_t h_t = \Delta h_t + \tilde{b}_t \cdot \nabla h_t + \varphi_t h_t\,. \tag{4.20}$$

This gives the Feynman–Kac representation

$$h_t(x) = \mathbb{E}\left[h_0\big(X_t^{t,x}\big) \exp\left(\int_0^t \varphi_s\big(X_s^{t,x}\big)\,\mathrm{d}s \right) \right],$$

where $X^{t,x}$ solves

$$X_0^{t,x} = x\,, \qquad \mathrm{d}X_s^{t,x} = \tilde{b}_{t-s}\big(X_s^{t,x}\big)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}B_t \quad \text{for } s \in [0,t].$$

Suppose additionally that φ_t, h_0 and $1/h_0$ are bounded and Lipschitz continuous (the general case being obtained afterwards by an approximation argument, which we omit here, referring to Chapter 3). Then, applying synchronous coupling to the Feynman–Kac formula above, for any $T > 0$ we obtain a constant $M > 0$ such that for every t, $s \in [0, T]$ and every x, $y \in \mathbb{R}^d$,

$$M^{-1} \leqslant h(t,x) \leqslant M \qquad \text{and} \qquad |h(t,x) - h(s,y)| \leqslant M\big(|t-s|^{1/2} + |x-y|\big)\,.$$

Taking the logarithm we obtain that $u_t := \ln h_t$ is a bounded and uniformly continuous viscosity solution to the HJB equation,

$$\partial_t u_t = \Delta u_t + |\nabla u_t|^2 + \tilde{b}_t \cdot \nabla u_t + \varphi_t\,. \tag{4.21}$$

In order to use a stochastic control representation of the solutions of such equations, for $N \in \mathbb{N}$, consider the approximative HJB equation,

$$u_0^N = u_0\,, \qquad \partial_t u_t^N = \Delta u_t^N + \sup_{\alpha:|\alpha|\leqslant N} \{2\alpha \cdot \nabla u_t^N - |\alpha|^2\} + \tilde{b} \cdot \nabla u^N + \varphi_t\,, \tag{4.22}$$

and the associated control problem,

$$V^N(T,x) = \sup_{\nu} \sup_{\alpha:|\alpha_t|\leqslant N} \mathbb{E}\left[u_0\big(X_T^{\alpha,x}\big) + \int_0^T \Big(\varphi_t\big(X_t^{\alpha,x}\big) - |\alpha_t|^2\Big)\,\mathrm{d}t\right], \tag{4.23}$$

where $\nu = \big(\Omega, F, (\mathcal{F}_\cdot), \mathbb{P}, (B_\cdot)\big)$ stands for a filter probability space with the usual conditions and an $(\mathcal{F}_\cdot)$-Brownian motion B, α is an $\mathbb{R}^d$-valued progressively measurable process such that $\int_0^T \mathbb{E}\big[|\alpha_t|^m\big]\,\mathrm{d}t$ is finite for every $m \in \mathbb{N}$, and $X^{\alpha,x}$ solves

$$X_0^{\alpha,x} = x\,, \qquad \mathrm{d}X_t^{\alpha,x} = \Big(\tilde{b}\big(X_t^{\alpha,x}\big) + 2\alpha_t\Big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,. \tag{4.24}$$

By Theorem IV.7.1 and the results in Sections V.3 and V.9 of [91], the value function V^N defined by (4.23) is a bounded and uniformly continuous viscosity solution to (4.22).

Suppose $u_0 = \ln(m_0/\mu_0)$ is the sum of an M^{u_0}-bounded and an L^{u_0}-Lipschitz function. As shown in Lemma 3.14, using a reflection coupling of two solutions of (4.24) with different initial conditions but using the same control α, we get that there exist C', $\kappa > 0$, depending only on ρ, L, R, such that for every x, $y \in \mathbb{R}^d$, $N \in \mathbb{N}$, $T > 0$ and $t > 0$, we have

$$\begin{aligned} |V^N(T,x) - V^N(T,y)| &\leqslant 2M^\varphi t + 2M^{u_0}\mathbb{1}_{T<1} \\ &\quad + C'\bigg(\mathbb{1}_{t<T}\frac{M^\varphi}{t} + L^\varphi + e^{-\kappa T}\big(L^{u_0} + \mathbb{1}_{T\geqslant 1}M^{u_0}\big)\bigg)|x-y|\,. \end{aligned} \tag{4.25}$$

We simply take $t = 1$. Since both u and V^N are bounded and uniformly continuous on $[0,T] \times \mathbb{R}^d$, we can apply the parabolic comparison for viscosity solutions on the whole space [72, Theorem 1] to obtain $V^N(T,x) = u_T(x)$ for N sufficiently large. Therefore, we have obtained that there exists $C > 0$ such that for every $T > 0$ and every x, $y \in \mathbb{R}^d$, we have

$$|u_T(x) - u_T(y)| \leqslant C(1 + |x-y|)\,. \tag{4.26}$$

Besides, in view of (4.25), we can find $C>0$ independent from m_0 such that (4.26) holds with this C for all T large enough. Moreover this C can be taken arbitrarily small provided $M^\varphi + L^\varphi$ is small enough.

We can then decompose u_T as the sum of a bounded and a Lipschitz continuous functions (with time uniform bounds for both functions). For instance we can consider a $2C(1+\sqrt{d})$-Lipschitz function v_T that coincides with u_T at all points $x\in\mathbb{Z}^d$ (thanks to (4.26)) and then $u_T - v_T$ is uniformly bounded (thanks to (4.26) again) uniformly in T. The proof is concluded by applying successively the Holley–Stroock and Aida–Shigekawa perturbation lemmas [113, 1]. $\square$

4.3 Sharp PoC for McKean–Vlasov diffusions

4.3.1 Settings and notations

In this section, we consider the non-linear McKean–Vlasov equation on $\mathbb{R}^d$:

$$\partial_t m_t = \nabla\cdot\left(\sigma^2\nabla m_t - F(\cdot,m_t)m_t\right), \tag{4.27}$$

which corresponds to (4.1) in the case (4.3). In fact, since we want to apply the results of [142], we consider its settings, which reads

$$F(x,m) = b_0(x) + \int_{\mathbb{R}^d} b(x,y)m(\mathrm{d}y)$$

for some $b_0:\mathbb{R}^d\to\mathbb{R}^d$ and $b:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d$ (which additionally may depend on time in [142], which we don't consider here for simplicity as it is not the case in the examples were are interested in, although it would work similarly). It is associated to the system of interacting particles $\boldsymbol{X}=(X^1,\dots,X^N)$ solving

$$\forall i\in[\![1,N]\!],\qquad \mathrm{d}X_t^i = b_0\big(X_t^i\big)\,\mathrm{d}t + \frac{1}{N-1}\sum_{j\in[\![1,N]\!]\setminus\{i\}} b\big(X_t^i,X_t^j\big)\,\mathrm{d}t + \sqrt{2}\sigma\,\mathrm{d}B_t^i\,, \tag{4.28}$$

where B^1, ..., B^N are independent d-dimensional Brownian motions. Denote by m_t^N the law of $\big(X_t^1,\dots,X_t^N\big)$ and by $m_t^{k,N}$ the law of $\big(X_t^1,\dots,X_t^k\big)$ for $k\leqslant N$.

The PoC phenomenon describes the fact that, in the system of interacting particles, as $N\to\infty$, particles become more and more independent, so that $m_t^{k,N}$ converges to $m_t^{\otimes k}$ for a fixed k. Up to recently, known results were typically that, under suitable conditions, for a fixed $t>0$, $\big\|m_t^{k,N} - m_t^{\otimes k}\big\|_{\mathrm{TV}} = \mathcal{O}\big(\sqrt{k/N}\big)$. This can be for instance obtained by showing the global estimate $\mathcal{H}\big(m_t^N\big|m_t^{\otimes N}\big)=\mathcal{O}(1)$ (which is optimal) using then that $\mathcal{H}\big(m_t^N\big|m_t^{\otimes N}\big) = (N/k)\mathcal{H}\big(m_t^{k,N}\big|m_t^{\otimes k}\big)$ (assuming for simplicity that $n/k\in\mathbb{N}$) and concluding with Pinsker's inequality. This k/N rate for the marginal relative entropy (hence $\sqrt{k/N}$ in TV) was thought to be optimal until Lacker showed in [140] that it is possible to get a rate k^2/N^2 by working with a BBGKY hierarchy of entropic bounds instead of simply with the full entropy of the N particles system. We refer to such entropic estimates with a rate k^2/N^2 as *sharp* PoC, by comparison with other results (the k^2/N^2 rate being optimal, as it is reached, e.g., in Gaussian cases). The work [140] deals with finite-time intervals, and the technique is then refined by Lacker and Le Flem in [142] to get uniform-in-time sharp PoC in some cases (small interaction in the torus or convex potentials in $\mathbb{R}^d$). A crucial ingredient in their result is a uniform LSI for the solution of the non-linear equation (4.27). Our results can thus be applied to extend their results to more general cases, allowing for instance for non-convex potentials on $\mathbb{R}^d$.

The rest of this section is organized as follows. In Section 4.3.2 for the reader's convenience we give a brief overview of the general result of Lacker and Le Flem. In Sections 4.3.3 and 4.3.4 we apply respectively Theorems 4.4 and 4.3 to get, under suitable conditions, uniform-in-time LSI for solutions of the McKean–Vlasov equation, and thus uniform-in-time sharp PoC as a corollary, in cases which are not covered by [142].

4.3.2 Lacker and Le Flem's result

First, for the reader's convenience, we recall [142, Theorem 2.1]. There are two sets of assumptions to apply this result: Assumption **E** of [142] is technical conditions related to well-posedness of m and m^N and we omit them as they are not important in our discussion (see Proposition 4.9 below). The second set of assumptions of [142] is the following.

Assumption 4.5 (Assumption **A** of [142])**.** The following holds.

1. $(m_t)_{t\geqslant 0}$ satisfies a uniform LSI with constant $\eta > 0$.
2. $(m_t)_{t\geqslant 0}$ satisfies a uniform transport inequality: there exists $\gamma > 0$ such that, for all $t \geqslant 0$, $x \in \mathbb{R}^d$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$,
$$\left|\nu\big(b(x,\cdot)\big) - m_t\big(b(x,\cdot)\big)\right|^2 \leqslant \gamma\mathcal{H}(\nu|m_t)\,. \tag{4.29}$$
3. $(m_t)_{t\geqslant 0}$ and $\big(m_t^N\big)_{t\geqslant 0}$ satisfy this uniform L^2 boundedness:
$$\sup_{N\in\mathbb{N}}\sup_{t\geqslant 0}\int_{\mathbb{R}^{dN}}\left|b(x_1,x_2) - m_t\big(b(x_1,\cdot)\big)\right|^2 m_t^N(\mathrm{d}x) < \infty\,. \tag{4.30}$$

When b is bounded, (4.30) is trivial and (4.29) follows from Pinsker's inequality. When $y \mapsto b(x,y)$ is Lipschitz continuous uniformly in x, (4.30) follows from time-uniform second moment bounds, which are classically obtained by Lyapunov arguments, and (4.29) is implied by the uniform LSI.

Theorem 4.6 (From Theorem 2.1 of [142])**.** *Under Assumptions **A** and **E** of [142], assume moreover that $\sigma^4 > 8\gamma\eta$ and that*

$$\exists C_0 > 0,\ \forall N \geqslant 2,\ \forall k \in [\![1,N]\!],\qquad \mathcal{H}\big(m_0^{k,N}\big|m_0^{\otimes k}\big) \leqslant C_0\frac{k^2}{N^2}\,. \tag{4.31}$$

Then,

$$\exists C > 0,\ \forall N \geqslant 2,\ \forall k \in [\![1,N]\!],\ \forall t \geqslant 0,\qquad \mathcal{H}\big(m_t^{k,N}\big|m_t^{\otimes k}\big) \leqslant C\frac{k^2}{N^2}\,. \tag{4.32}$$

Remark 4.7. As in Remark 4.2, it is in fact sufficient to enforce Assumption **A** with the condition $\sigma^4 > 8\gamma\eta$ for times $t \geqslant t_0$ for some t_0 and apply Theorem 4.6 to $(m_{t+t_0})_{t\geqslant 0}$. More precisely, for some t_0, assume that (4.29) and (4.30) holds uniformly over $t \in [0,t_0]$. Then, assuming the initial chaos (4.31), [140, Theorem 2.2] gives (4.32) for some constant $C > 0$ uniformly over $t \in [0,t_0]$. In particular, the initial chaos (4.31) holds for $(m_{t+t_0})_{t\geqslant 0}$.

In [142], the assumptions of Theorem 4.6 are shown to hold in two cases: either convex potentials on $\mathbb{R}^d$, or models on the torus. In any cases, the condition $\sigma^4 > 8\gamma\eta$ (corresponding to $r_c > 1$ with the notation of [142, Theorem 2.1]) means that the PoC estimates require that either the temperature σ^2 is high enough or the strength of the interaction is small enough. In Sections 4.3.3 and 4.3.4 we extend the range of application of [142, Theorem 2.1] to some cases with non-convex potentials on $\mathbb{R}^d$.

Before that, in order to focus on the uniform LSI afterwards, let us state a result concerning the other technical conditions, which is sufficient for the cases considered in the two next sections.

Assumption 4.8. The initial conditions m_0 and m_0^N have finite moments of all orders, m_0^N is exchangeable and there exists C independent from N such that $\int_{\mathbb{R}^d} |x_1|^2 m_0^{1,N}(\mathrm{d}x_1) \leqslant C$.

We omit the proof of the next result, the arguments are the same as in [142, Corollary 2.7].

Proposition 4.9. *Assume that b_0 are b are $\mathcal{C}^1$, that $|b_0|$ grows at most polynomially, that b is the sum of a bounded and a Lipchitz continuous function, and that there exists c, $C > 0$ such that for all x, $y \in \mathbb{R}^d$,*

$$\big(b_0(x) + b(x,y)\big) \cdot x \leqslant -c|x|^2 + C(1 + |y|)\,.$$

*Then, under Assumption 4.8, $(m_t)_{t\geqslant 0}$ and $\big(m_t^N\big)_{t\geqslant 0}$ are well defined and Assumption **E** of [142] and the uniform L^2 boundedness (4.30) holds.*

4.3.3 Convergent trajectories

In this section we focus on the cases where m_t converges as $t \to \infty$ towards a stationary solution m_* of the non-linear equation (4.27). This is known to hold in various cases of interest, like the granular media equation with convex potentials, or repulsive interaction, or high temperature, or small interaction, or other models like the adaptive biasing force method [150] or the mean-field gradient descent ascent [156]. So, assume that

$$\|m_t - m_*\|_{\mathrm{TV}} \underset{t\to\infty}{\longrightarrow} 0\,. \tag{4.33}$$

Remark 4.10. Under suitable conditions, [192, Theorem 4.1] allows to obtain (4.33) from a $\mathcal{W}_2$ convergence.

We now discuss suitable conditions to apply Theorem 4.4 with M^φ, L^φ arbitrarily small for large times, where we decompose the drift $F(x,m_t) = a_0(x) + g_t(x)$ with $a_0(x) = F(x,m_*)$ and $g_t(x) = F(x,m_t) - F(x,m_*)$. For simplicity we focus on the case where

$$F(x,m) = -\nabla V(x) - \int_{\mathbb{R}^d} \nabla_x W(x,y) m(\mathrm{d}y)\,, \tag{4.34}$$

for some $V \in \mathcal{C}^2(\mathbb{R}^d,\mathbb{R})$ and $W \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d,\mathbb{R})$. The next result would be easily adapted to other cases where the density of the stationary solutions of (4.27) are explicit or solve an explicit fixed-point equation (namely when the invariant measure of $\sigma^2\Delta + F(\cdot,m)\cdot\nabla$ is explicit for each m), which is for instance the case in [150, 156].

Proposition 4.11. *Let $(m_t)_{t\geqslant 0}$ be a solution to (4.27) (in the case (4.34)) which converges in TV in long time towards a stationary solution m_*. Assume that m_0 admits a density e^{u_0} with respect to m_*, with u_0 being the sum of a bounded and a Lipschitz continuous function. Assume furthermore that there exists L, $\alpha > 0$ such that, for all x, $y \in \mathbb{R}^d$,*

$$|\Delta_x W(x,y)| \leqslant L\,, \qquad |\nabla_x W(x,y)| \leqslant \frac{L}{1+|x-y|^\alpha}\,, \qquad |\nabla V(x)| \leqslant L(1+|x|^\alpha)\,. \tag{4.35}$$

Finally, assume that V is strongly convex outside of a compact set. Then, $(m_t)_{t\geqslant 0}$ satisfies a uniform LSI. Moreover, as $t\to\infty$, the optimal LSI constant of m_t converges to the optimal LSI constant of m_*.

Notice that, V being strongly convex outside a compact set, the last condition of (4.35) can only hold with some $\alpha \geqslant 1$. Hence, the second condition of (4.35) on ∇W means that we only consider local interactions.

Proof. Considering the decomposition $F(x,m_t) = a_0(x) + g_t(x)$ with $a_0(x) = F(x,m_*)$ and $g_t(x) = F(x,m_t) - F(x,m_*)$, we have to show that Theorem 4.4 applies to $(m_{t+t_0})_{t\geqslant 0}$ with M^φ, L^φ arbitrarily small provided t_0 is large enough. Indeed, the last part of Theorem 4.4 will then give that, for any $\varepsilon > 0$, the optimal LSI constant of m_t is less than $C_0 + \varepsilon$ for t large enough, where C_0 is the optimal LSI constant of m_*. On the other hand, for any $\varepsilon > 0$, there exists a non-constant $\mathcal{C}^\infty$ function f with compact support such that

$$m_*(f^2\ln f^2) - m_*(f^2)\ln m_*(f^2) \geqslant (C_0-\varepsilon) m_*|\nabla f|^2\,.$$

The weak convergence implied by (4.33) leads to

$$m_t(f^2\ln f^2) - m_t(f^2)\ln m_t(f^2) \geqslant (C_0-2\varepsilon) m_t|\nabla f|^2$$

for t large enough, which implies that the optimal LSI constant of m_t is larger than $C_0 - 2\varepsilon$.

Hence, we turn to the application of Theorem 4.4 using its notations. We write $m\star W(x) = \int_{\mathbb{R}^d} W(x,y)m(\mathrm{d}y)$. The invariant measure of $a_0\cdot\nabla + \sigma^2\Delta$ is $\mu_0 = m_*$, with $\nabla \ln m_* = -\nabla(V + m_*\star W) = F(\cdot, m_*)$, so that

$$\tilde{b}_t(x) = -\nabla(V + 2m_*\star W - m_t\star W)\,.$$

Since $\nabla_x W$ is bounded by (4.35), the contribution of W in $\tilde{b}_t$ is bounded (uniformly in t) and thus (4.9) holds thanks to the convexity of V outside a compact set. From (4.35),

$$|\nabla\cdot g_t(x)| = |(m_t - m_*)\star \Delta_x W(x)| \leqslant L\|m_t - m_*\|_{\mathrm{TV}}\,,$$

and, given (Y,Y') an optimal TV coupling of m_t and m_* and using the Cauchy–Schwarz inequality,

$$\begin{aligned}
&|g_t(x)\cdot\nabla\ln m_*(x)| \\
&\quad\leqslant \left|\mathbb{E}[\nabla_x W(x,Y) - \nabla_x W(x,Y')]\right| L(2+|x|^\alpha) \\
&\quad\leqslant \mathbb{E}\left[\mathbb{1}_{Y\neq Y'}\left(\frac{1}{1+|x-Y|^\alpha} + \frac{1}{1+|x-Y'|^\alpha}\right)\right] L^2(2+|x|^\alpha) \\
&\quad\leqslant \|m_t - m_*\|_{\mathrm{TV}}^{1/2}\,\mathbb{E}\left[\left(\frac{1}{1+|x-Y|^\alpha} + \frac{1}{1+|x-Y'|^\alpha}\right)^2\right]^{1/2} L^2(2+|x|^\alpha)\,.
\end{aligned}$$

Then we bound, for the first term in the expection,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{(1+|x-Y|^{\alpha})^2}\right] &\leqslant \frac{1}{(1+|x/2|^{\alpha})^2} + \mathbb{P}[|Y|\geqslant |x|/2]\\ &\leqslant \frac{1}{(1+|x/2|^{\alpha})^2} + \frac{1+\mathbb{E}[|Y|^{2\alpha}]}{1+|x/2|^{2\alpha}}\,,\end{aligned}$$

and similarly for the second term involving m_*. Using that V is convex outside a compact set and that $\nabla_x W$ is bounded we easily get by Lyapunov arguments that the moments of m_t are bounded uniformly in time. As a consequence, we have obtained, for $\varphi_t := -\nabla\cdot g_t + g_t\cdot\nabla\ln\mu_0$, a bound

$$\|\varphi_t\|_\infty \leqslant L'\|m_t - m_*\|_{\mathrm{TV}}^{1/2}$$

for some L' independent from t. The TV convergence (4.33) concludes the proof. □

Corollary 4.12. *Under Assumption 4.8 and the settings of Proposition 4.11, assume furthermore that W is bounded and $V = V_1 + V_2$ where V_1 is ρ-strongly convex and V_2 is bounded. Assume that*

$$\sigma^2 > \frac{4}{\rho}\|\nabla_x W\|_\infty^2 \exp\left(\frac{\|V_2\|_\infty + \|W\|_\infty}{\sigma^2}\right). \tag{4.36}$$

Then, provided the initial PoC (4.31) holds, so does the uniform in time sharp PoC (4.32).

This applies to cases on $\mathbb{R}^d$ where V is not convex, which are not covered by [142]. In general cases where V may have several local minima, a condition in the spirit of (4.36), that states that either temperature is large enough or interaction is small enough, is necessary to have a uniform-in-time propagation of chaos estimate.

Proof. The assumptions of Proposition 4.11 imply those of Proposition 4.9. Since $\nabla_x W$ is bounded, Pinsker's inequality gives the transport inequality (4.29) with $\gamma = \|\nabla_x W\|_\infty^2/2$. Proposition 4.11 provides the uniform LSI for $(m_t)_{t\geqslant 0}$. Moreover, for large times, the LSI constant of m_t converges to the LSI constant C_* of m_*, which by the Bakry–Émery and Holley–Stroock results is less than $\sigma^2\rho^{-1}\exp\big((\|V_2\|_\infty + \|W\|_\infty)/\sigma^2\big)$. Corollary 4.12 thus follows from Theorem 4.6 (since, as noticed in Remark 4.7, the condition $\sigma^4 > 8\gamma\eta$ only has to be verified for sufficiently long times). □

4.3.4 High temperature regime

Instead of Corollary 4.12, using Theorem 4.3, we can get an alternative result, which doesn't require the a priori knowledge that m_t converges in large time and with weaker assumptions on W, but which only works at high temperature and is less explicit (an explicit condition on σ^2 can be obtained in principle by checking the proofs, but it wouldn't be as nice as (4.36)). In the next statement we consider a solution $(m_t)_{t\geqslant 0}$ of (4.27) in the case (4.34).

Proposition 4.13. *Under Assumption 4.8, assume furthemore that $|\nabla U|$ grows at most polynomially, that there exist ρ, L, $R > 0$ such that, for all $z \in \mathbb{R}^d$, $\psi_z := -\nabla U - \nabla_x W(\cdot, z)$ satisfies*

$$\big(\psi_z(x) - \psi_z(y)\big) \cdot (x-y) \leqslant \begin{cases} -\rho|x-y|^2 & \forall x, y \in \mathbb{R}^d \text{ with } |x| \geqslant R\,, \\ L|x-y|^2 & \forall x, y \in \mathbb{R}^d\,, \end{cases} \tag{4.37}$$

and that $\nabla_x W = F_1 + F_2$ where F_1 is bounded and $y \mapsto F_2(x,y)$ is L_W-Lipschitz with $8L_W^2 < \rho$, uniformly in x. Then, there exists $\sigma_*^2 > 0$ (which depends on U, W and d) such that, assuming $\sigma^2 \geqslant \sigma_*^2$ and the initial sharp PoC (4.31), we have that the uniform in time sharp PoC (4.32) holds.

In particular, if U is strongly convex outside a compact set and $x \mapsto W(x,z)$ is convex for all z with $\nabla_x W$ being bounded (e.g. $W(x,z) = a\sqrt{1+|x-z|^2}$ for $a > 0$), then Proposition 4.13 applies without requiring the interaction to be small (although the temperature threshold σ_*^2 can become large when the interaction is strong).

Proof. We verify the conditions of Theorem 4.6. Using (4.37) with $y = 0$ we see that Proposition 4.9 holds. The uniform LSI in the high temperature regime $\sigma^2 \geqslant \sigma_0^2$ is ensured by Theorem 4.3, and for times large enough it holds with a constant $\eta = \sigma^2\eta'$ for some $\eta' > 0$ independent from σ, and which can be taken arbitrarily close to $1/\rho$ for σ^2 large enough. Here we have used that $\sup\{-x \cdot b_t(x) : |x| \leqslant R_*\}$ can be bounded by a constant K independent from t and such that (4.8) holds for σ large enough (for t large enough). Indeed, we can bound

$$|b_t(x)| \leqslant |\nabla U(x)| + \|F_1\|_\infty + |F_2(x,0)| + L_W \int_{\mathbb{R}^d} |y| m_t(\mathrm{d}y)\,.$$

Then, the condition (4.37) implies that $s_t := \int_{\mathbb{R}^d} |y|^2 m_t(\mathrm{d}y)$ satisfies $\mathrm{d}s_t/\,\mathrm{d}t \leqslant -\rho s_t/2 + q + 2d\sigma^2$ for some $q > 0$ independent from t and σ^2. From this, for t large enough, we get $\int_{\mathbb{R}^d} |y| m_t(\mathrm{d}y) \leqslant C(1+\sigma)$ where C depends only on d, ρ, L, R. As a consequence, in (4.7) we can take $K = C'(1+\sigma)$ for some C' (independent from t and σ), so that (4.8) holds for σ large enough, as claimed.

It remains to check the transport inequality (4.29). For any $\theta > 0$ we can bound, for all $t \geqslant 0$, $x \in \mathbb{R}^d$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$,

$$\begin{aligned} &\big|\nu\big(b(x,\cdot)\big) - m_t\big(b(x,\cdot)\big)\big|^2 \\ &\quad\leqslant (1+\theta)\big|\nu\big(F_1(x,\cdot)\big) - m_t\big(F_1(x,\cdot)\big)\big|^2 + (1+\theta^{-1})\big|\nu\big(F_2(x,\cdot)\big) - m_t\big(F_2(x,\cdot)\big)\big|^2 \\ &\quad\leqslant (1+\theta)\|F_1\|_\infty^2 \|\nu - m_t\|_{\mathrm{TV}}^2 + (1+\theta^{-1}) L_W^2 \mathcal{W}_2^2(\nu, m_t) \\ &\quad\leqslant \gamma \mathcal{H}(\nu|m_t)\,, \end{aligned}$$

where we used Pinsker's and Talagrand's inequalities, and γ on the last line is defined by

$$\gamma = \frac{1+\theta}{2}\|F_1\|_\infty^2 + \sigma^2\eta'(1+\theta^{-1})L_W^2\,.$$

Fixing θ (independent from σ) large enough so that $8(1+\theta^{-1})L_W^2 < \rho$, the condition $\sigma^4 > 8\gamma\eta$ holds for σ large enough, which concludes. □

4.4 Application to log and Riesz interactions

In this section, we still consider McKean–Vlasov equations (4.27), but now we impose the following condition on the non-linear drift.

Assumption 4.14. We have $d \geqslant 2$, $s \in [0, d-1)$ and the McKean–Vlasov drift in (4.27) reads

$$F(x,m) = -\nabla U(x) + M\nabla g_s \star m(x)\,,$$

where U, M, g_s satisfy the following conditions:

- the function $U : \mathbb{R}^d \to \mathbb{R}$ has bounded Hessian $\nabla^2 U \in L^\infty$ and satisfies the weak convexity condition: there exist $\kappa_U > 0$ and $R \geqslant 0$ such that for all x, $y \in \mathbb{R}^d$ with $|x-y| \geqslant R$, we have

$$\langle \nabla U(x) - \nabla U(y), x-y \rangle \geqslant \kappa_U |x-y|^2\,;$$

- $g_s : \mathbb{R}^d \to \mathbb{R}$ is the logarithmic or Riesz potential:

$$g_s(x) = \begin{cases} -\ln|x| & \text{when } s = 0, \\ |x|^{-s} & \text{when } s > 0; \end{cases}$$

- in the sub-Coulombic case where $s < d-2$, M is a $d \times d$ real matrix such that $M : \nabla^2 g(x) \geqslant 0$ for $x \neq 0$; in the Coulombic and the super-Coulombic cases where $s \in [d-2, d-1)$, M is anti-symmetric.

These models have raised a high interest over the recent years, in particular with a series of work by Rosenzweig, Serfaty and coauthors on the one hand (see e.g. [199, 59, 200] and references within) and Bresch, Jabin, Wang and coauthors on the other hand (see e.g. [124, 32, 31] and references within). The main result of the section, to be stated in Theorem 4.25 in Section 4.4.4, addresses the McKean–Vlasov drift force above with $d \geqslant 2$, $s = 0$, M being anti-symmetric and U being isotropically quadratic. We show that in this case the dynamics exhibits the time-uniform propagation of chaos. This result is a continuation of a recent work of Guillin, Le Bris and one of the author [98], where the uniform PoC is shown for the dynamics on the torus (thus in a periodic setting). We also note that a non-time-uniform result on the whole space have also been obtained very recently by Feng and Wang [90]. In terms of methodology, the main addition of our work is that we employ the reflection coupling technique of Conforti [61] to get regularity bounds for the mean field flow on the whole space (Theorems 4.19 and 4.24), which enable to apply the Jabin–Wang method.

We will write $g = g_s$ if that does not lead to ambiguities. For simplicity, we also set $\sigma = 1$ in this section. Under the assumptions above, we denote $K = M\nabla g$, and the McKean–Vlasov dynamics writes

$$\partial_t m_t = \Delta m_t - \nabla \cdot \big(m_t (K \star m_t - \nabla U) \big)\,. \tag{4.38}$$

Consider now the system of N particles in interaction:

$$\mathrm{d}X_t^i = -\nabla U\big(X_t^i\big)\,\mathrm{d}t + \frac{1}{N-1} \sum_{j \in [\![1,N]\!] \setminus \{i\}} K\big(X_t^i - X_t^j\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i\,, \quad i = 1, \ldots, N, \tag{4.39}$$

where W_t^i are N independent Brownian motions. The flow $m_t^N = \mathrm{Law}(\boldsymbol{X}_t) = \mathrm{Law}(X_t^1, \dots, X_t^N)$ of probabilities in $\mathbb{R}^{dN}$ satisfies the Fokker–Planck equation at least formally:

$$\partial_t m_t^N = \sum_{i=1}^N \biggl(\Delta_i m_t^N - \nabla_i \cdot \biggl(\Bigl(\frac{1}{N-1} \sum_{j \in [\![1,N]\!] \setminus \{i\}} K(x_i - x_j) - \nabla U(x^i) \Bigr) m_t^N \biggr) \biggr) . \tag{4.40}$$

In this section, η^ε denotes a $\mathcal{C}^\infty$ mollifier with support in $B(0,\varepsilon)$ that is also invariant by rotation. We set $g^\varepsilon := g \star \eta^\varepsilon$ and $K^\varepsilon := M\nabla g^\varepsilon = M\nabla g \star \eta^\varepsilon$. Since under Assumption 4.14, we are restricted to the case where $s < d-1$, the interaction potential $g \propto |x|^{-s}$ is integrable around zero, so g^ε is infinitely differentiable with bounded derivatives. Notice that the rotational invariance of η^ε implies that the value $g^\varepsilon(x)$ depends only on $|x|$ and thus, $\nabla g^\varepsilon(x)$ is parallel to x. We also work with the approximation of the confinement $U^\varepsilon := U \star \eta^\varepsilon$.

Sometimes, in the rest of this section, for conciseness, we write $A \lesssim B$ when there exists a constant C such that $A \leqslant CB$.

4.4.1 Well-posedness of the mean field and particle systems

For a function $f : \mathbb{R}^d \to \mathbb{R}$ and $\theta \in (0,1]$, we denote the homogeneous θ-Hölder (semi-)norm of f by

$$[f]_{\mathcal{C}^\theta} = \sup_{x,y \in \mathbb{R}^d \,:\, x \neq y} \frac{|f(x) - f(y)|}{|x-y|^\theta} .$$

In order to study the singular interaction kernel K, we use the following crucial estimate. This generalizes the estimate in (2.9) of [199] (which corresponds to the case $p = \infty$). We refer readers to Lemma 4.5.4 and Theorem 4.5.10 of [114] for the proof, where the statement of the latter should be accompanied with an interpolation.

Proposition 4.15. *Let $s > 0$. For all $m \in L^1 \cap L^p(\mathbb{R}^d)$ with $\bigl(1 - \frac{s}{d}\bigr)^{-1} < p \leqslant \infty$, we have*

$$\bigl\| |\cdot|^{-s} \star m \bigr\|_{L^\infty} \lesssim \|m\|_{L^1}^{1-qs/d} \|m\|_{L^p}^{qs/d} ,$$

where $p^{-1} + q^{-1} = 1$. If additionally, for some $\theta \in (0,1)$, we have $\bigl(1 - \frac{s+\theta}{d}\bigr)^{-1} < p \leqslant \infty$, then

$$\bigl[|\cdot|^{-s} \star m \bigr]_{\mathcal{C}^\theta} \lesssim \|m\|_{L^1}^{1-q(s+\theta)/d} \|m\|_{L^p}^{q(s+\theta)/d} .$$

Then, we present the well-posedness results for the mean field and the particle system.

Proposition 4.16 (Well-posedness of the mean field system). *Let Assumption 4.14 hold. Then we have the following results:*

- *For each initial value $m_0 \in L^1 \cap L^\infty \cap \mathcal{P}(\mathbb{R}^d)$, there exists a unique solution to the mean field flow (4.38) in $C\bigl([0,\infty); L^1(\mathbb{R}^d) \cap \mathcal{P}\bigr) \cap L^\infty\bigl([0,\infty); L^\infty(\mathbb{R}^d)\bigr)$ depending continuously on the initial value. In particular, we have the time-uniform bound:*

$$\sup_{t \in [0,\infty)} \|m_t\|_{L^\infty} \leqslant C_1(U, \|m_0\|_{L^\infty}) < \infty . \tag{4.41}$$

- *If additionally the initial value m_0 has finite k-th moment for some $k > 0$, then the mean field flow m_t has finite k-th moment, uniformly in time:*

$$\sup_{t\in[0,\infty)} \int_{\mathbb{R}^d} |x|^k m_t(\mathrm{d}x) \leqslant C_2\biggl(U, K, k, \|m_0\|_{L^\infty}, \int_{\mathbb{R}^d} |x|^k m_0(\mathrm{d}x)\biggr)$$

- *Finally, let $K^\varepsilon = K \star \eta^\varepsilon$, $U^\varepsilon = U \star \eta^\varepsilon$ be the mollified kernel and confinement. If m_0^ε converges to m_0 in L^1 and if $\sup_\varepsilon \|m_0^\varepsilon\|_{L^\infty} < \infty$, then the solution m_t^ε of the approximate mean field flow*

$$\partial_t m_t^\varepsilon = \Delta m_t^\varepsilon - \nabla \cdot \bigl(m_t^\varepsilon (K^\varepsilon \star m_t^\varepsilon - \nabla U^\varepsilon)\bigr) \tag{4.42}$$

converges to m_t in L^1 for all $t \geqslant 0$. Moreover, the L^∞ norm and the k-th moment bounds above hold when we replace m by m^ε.

Proposition 4.17 (Well-posedness of the particle system)**.** *Let Assumption 4.14 hold with $s \leqslant d-2$ and suppose that for all $x \in \mathbb{R}^d$, we have $x^\top M x \leqslant 0$. Then, for any initial value $\boldsymbol{X}_0 = \bigl(X_0^1, \dots, X_0^N\bigr)$ such that $X_0^i \neq X_0^j$ almost surely for $i \neq j$, the SDE system (4.39) has a global unique strong solution. Moreover, setting $K^\varepsilon = K \star \eta^\varepsilon$, $U^\varepsilon = U \star \eta^\varepsilon$, and considering the approximate SDE system*

$$\mathrm{d}X_t^{\varepsilon,i} = -\nabla U^\varepsilon\bigl(X_t^{\varepsilon,i}\bigr)\,\mathrm{d}t + \frac{1}{N-1} \sum_{j\in[\![1,N]\!]\setminus\{i\}} K^\varepsilon\bigl(X_t^{\varepsilon,i} - X_t^{\varepsilon,j}\bigr)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i\,, \tag{4.43}$$

for $i \in [\![1,N]\!]$, with the initial condition $X_0^{\varepsilon,i} = X_0^i$, we have, for all $t \geqslant 0$ and $i = 1, \dots, N$,

$$X_t^{\varepsilon,i} \to X_t^i \text{ a.s.,} \qquad \text{when } \varepsilon \to 0\,.$$

These results may be considered mathematical folklore and we do not claim originality from them. Their proofs are postponed to Appendix C.1.

4.4.2 Uniform Lipschitz and Hessian bounds, and LSI

We introduce the invariant measure μ_0 of the reversible diffusion generated by $\Delta - \nabla U \cdot \nabla$, whose density is explicit:

$$\mu_0(x) = Z(\mu_0)^{-1} \exp\bigl(-U(x)\bigr)\,, \quad Z(\mu_0) = \int_{\mathbb{R}^d} \exp\bigl(-U(x)\bigr)\,\mathrm{d}x\,.$$

Note that, under Assumption 4.14, using the HJB flow method of Conforti (see Theorem 1.3 and Remark 1.7 of [62]), we can show that the measure μ_0 is the image of a Gaussian measure under a transport mapping with an explicit Lipschitz constant, and thus satisfies an LSI with an explicit constant.

We use the following result on the Lipschitz and Hessian bounds on the solution to a class of HJB equations.

Theorem 4.18. *Let $T > 0$. Suppose that $\tilde b \in \mathcal{C}^{0,\infty}([0,T]\times\mathbb{R}^d;\mathbb{R}^d)$, $\varphi \in \mathcal{C}^{0,\infty}([0,T]\times\mathbb{R}^d;\mathbb{R})$, $u_0 \in \mathcal{C}^\infty(\mathbb{R}^d;\mathbb{R})$, and their space derivatives $\nabla^k \tilde b$, $\nabla^k \varphi$, $\nabla^k u_0$ are bounded for all $k \geqslant 1$. Then there exists a spatially $\mathcal{C}^2$ solution u to the HJB equation*

$$\partial_t u_t = \Delta u_t - |\nabla u_t|^2 + \tilde b_t \cdot \nabla u_t + \varphi_t$$

and this solution is unique within the class of spatially Lipschitz functions.

Furthermore, suppose the drift $\tilde b$ satisfies the weak convexity condition

$$\big(\tilde b_t(x) - \tilde b_t(y), x - y\big) \leqslant \kappa_{\tilde b}(|x-y|)|x-y|$$

for some $\mathcal{C}^1$-continuous $\kappa_{\tilde b} : (0,\infty) \to \mathbb{R}$ such that $\int_0^1 r\big(\kappa_{\tilde b}(r) \vee 0\big)\,\mathrm{d}r < \infty$ and $\liminf_{r\to\infty} \kappa_{\tilde b}(r) < 0$. Then, we have the following quantitative estimates on u:

- *If $\varphi_t \in L^\infty$ for all $t \in [0,T]$, then, we have, for all $t \in [0,T]$,*

$$\|\nabla u_t\|_{L^\infty} \leqslant Ce^{-ct}\|\nabla u_0\|_{L^\infty} + \int_0^t \frac{Ce^{-cv}}{\sqrt{v\wedge 1}}\|\varphi_{t-v}\|_{L^\infty}\,\mathrm{d}v\,, \tag{4.44}$$

  where C, $c > 0$ and depend only on $\kappa_{\tilde b}$.

- *If additionally, $\nabla\varphi_t \in L^\infty$ for all $t \in [0,T]$, then we have, for all $t \in [0,T]$,*

$$\begin{aligned}\|\nabla^2 u_t\|_{L^\infty} \leqslant{} & \frac{C'e^{-c't}}{\sqrt{t\wedge 1}}\|\nabla u_0\|_{L^\infty} \\ & + \int_0^t \frac{C'e^{-c'v}}{\sqrt{v\wedge 1}}\big(\|\nabla\varphi_{t-v}\|_{L^\infty} + \|\nabla\tilde b_{t-v}\cdot\nabla u_{t-v}\|_{L^\infty}\big)\,\mathrm{d}v\,, \end{aligned}\tag{4.45}$$

  where C', $c' > 0$ and depend only on $\kappa_{\tilde b}$, $\|\nabla u_0\|_{L^\infty}$ and $\sup_{t\in[0,T]}\|\varphi_t\|_{L^\infty}$.

The theorem is only an enhancement to the result of Conforti [61] by using the short-time gradient estimates obtained by Priola and Wang [187], and by Porretta and Priola [186]. Thus we only provide a sketch of proof here.

Sketch of proof of Theorem 4.18. The existence and uniqueness of the classical solution follow from standard arguments; for details, we refer readers to [61, Proof of Proposition 3.1]. The quantitative results differ from the main result of [61], specifically Theorem 1.3 therein, in only two respects: first, our analysis is conducted in a time-inhomogeneous setting; second, the uniform gradient estimate employed in our proof exhibits a blow-up rate of $t^{-1/2}$, as opposed to t^{-1}, as $t \to 0$.

Following the method in the proof of Theorem 4.4 (and ignoring technical issues about the correspondance to stochastic control problems), for every x, $y \in \mathbb{R}^d$ and $t \in [0,T]$, we can find stochastic processes $X^{\alpha,x}_\cdot$, $X^{\alpha,y}_\cdot$, α, all defined on $[0,t]$ and taking values in $\mathbb{R}^d$, such that

$$\begin{aligned} X^{\alpha,z}_0 = z\,, \qquad \mathrm{d}X^{\alpha,z}_v &= \Big(\tilde b\big(X^{\alpha,z}_v\big) + 2\alpha_v\Big)\,\mathrm{d}v + \sqrt{2}\,\mathrm{d}B^z_v\,, \qquad \text{for } z = x, y\,, \\ u(t,x) &= \mathbb{E}\bigg[\int_0^t \Big(|\alpha_v|^2 + \varphi_{t-v}\big(X^{\alpha,x}_v\big)\Big)\,\mathrm{d}v + u\big(0, X^{\alpha,x}_t\big)\bigg]\,, \\ u(t,y) &\leqslant \mathbb{E}\bigg[\int_0^t \Big(|\alpha_v|^2 + \varphi_{t-v}\big(X^{\alpha,y}_v\big)\Big)\,\mathrm{d}v + u\big(0, X^{\alpha,y}_t\big)\bigg]\,, \end{aligned}$$

where B^x, B^y are Brownian motions coupled by reflection until $X^{\alpha,x}$, $X^{\alpha,y}$ collide:

$$\mathrm{d}B^{\alpha,y}_v = \left(1 - \frac{2\big(X^{\alpha,y}_v - X^{\alpha,x}_v\big)\big(X^{\alpha,y}_v - X^{\alpha,x}_v\big)^\top}{\big|X^{\alpha,y}_v - X^{\alpha,x}_v\big|^2}\right)\mathrm{d}B^{\alpha,x}_v\,,$$

for $v \leqslant \tau \coloneqq \inf\big\{w \geqslant 0 : X^{\alpha,x}_w = X^{\alpha,y}_w\big\}$, and $\mathrm{d}B^{\alpha,x}_v = \mathrm{d}B^{\alpha,y}_v$ for $v > \tau$. Then, by subtracting the dynamics of $X^{\alpha,x}$ and $X^{\alpha,y}$, we find that their difference process

$\left|X_\cdot^{\alpha,x} - X_\cdot^{\alpha,y}\right|$ is stochastically dominated by a one-dimensional Itō process $(r_t)_{t\geqslant 0}$ solving

$$\mathrm{d}r_v = -r_v\kappa_{\tilde{b}}(r_v)\,\mathrm{d}v + 2\sqrt{2}\,\mathrm{d}W_v$$

with an absorbing boundary at 0 with initial value $r_0 = |x-y|$. It is shown in [187] that

$$\mathbb{P}[r_v > 0] \leqslant \frac{Cr_0}{\sqrt{v\wedge 1}}$$

for some C depending only on $\kappa_{\tilde{b}}$. Then, by combining the result above with the long-time Wasserstein contraction studied in [83], we get, for all $v\in[0,t]$,

$$\mathbb{P}[r_v > 0] \leqslant \frac{C'e^{-c'v}r_0}{\sqrt{v\wedge 1}}$$

for some C', $c'>0$ depending only on $\kappa_{\tilde{b}}$. Therefore, by subtracting the stochastic representation for $u(t,x)$, $u(t,y)$ and applying the bound above on r_v, we get the first claim.

For the second-order estimate, we take spatial derivatives in the HJB and find that ∇u_t solves the $\mathbb{R}^d$-valued equation

$$\partial_t\nabla u_t = \Delta\nabla u_t + (\tilde{b}_t - 2\nabla u_t)\cdot\nabla^2 u_t + \nabla\tilde{b}_t\cdot\nabla u_t + \nabla\varphi_t\,.$$

Thus, ∇u_t solves a second-order equation with the weakly semi-monotone drift term $\tilde{b}_t - 2\nabla u_t$ (as $\tilde{b}_t$ is weakly semi-monotone and ∇u_t is bounded by the first claim), and a bounded source term $\nabla\tilde{b}_t\cdot\nabla u_t + \nabla\varphi_t$. Writing the Feynman–Kac formula for ∇u_t and using the coupling by reflection as above, we get the second claim. □

Theorem 4.19. *Let Assumption 4.14 hold. Let $m_0\in\mathcal{P}(\mathbb{R}^d)$ be such that*

$$u_0 \coloneqq -\ln\frac{\mathrm{d}m_0}{\mathrm{d}\mu_0} = -\ln m_0 - U - \ln Z(\mu_0)$$

is Lipschitz continuous and let $(m_t)_{t\geqslant 0}$ be the solution to (4.38). Denote $u_t \coloneqq -\ln\mathrm{d}m_t/\mathrm{d}\mu_0$. Then we have, for all $t>0$,

$$\sup_{x\in\mathbb{R}^d}\left|K\star m_t(x)(1+|x|)\right| \leqslant C\,,\quad \|\nabla u_t\|_{L^\infty}\leqslant C\,,\quad \|\nabla^2 u_t\|_{L^\infty}\leqslant\frac{C}{\sqrt{t\wedge 1}}$$

for some C depending only on d, s, U, $|M|$, $\|m_0\|_{L^\infty}$ and $\|\nabla u_0\|_{L^\infty}$. Moreover, when $|M|$ increases and all other dependencies are kept constant, C increases. Consequently, the flow $(m_t)_{t\geqslant 0}$ satisfies a uniform LSI whose constant has the same dependency as above and is increasing in $|M|$.

The proof of Theorem 4.19 is postponed to Section 4.4.5.

Remark 4.20 (Modulated free energy and LSI, and kinetic case). We remark that since we have obtained the L^∞ bound of $\nabla^2 u_t$ in the theorem above (and also in Theorem 4.24 below), we can control the Lipschitz norm of the time-dependent vector field

$$x\mapsto\sigma^2\nabla\ln\frac{m_t(x)}{e^{-U(x)}} - K\star m_t(x). \tag{4.46}$$

The control of this quantity, as remarked in [59, Section 1.2], is crucial for the modulated free energy method since it appears in the "commutator estimates". See e.g. [209, Proposition 1.1] or [59, Proposition 2.13]. We note that unfortunately our method to obtain this control exploits the long-time contractivity of Brownian motions coupled by reflection, and relies fundamentally on the diffusivity of the dynamics, so it is not useful for deterministic dynamics (i.e. $\sigma = 0$) considered originally in [209]. Nevertheless, since similar results for the kinetic case in the time-homogeneous setting have been established by two of the authors in Theorems 3.6 and 3.20, our method provides bounds on $\nabla^2 \ln m_t$, which is of interest in the perspective of applying the arguments [124, Theorem 2] in such hypoelliptic cases.

Besides, together with the control of the Lipschitz norm of (4.46), a key ingredient to get uniform-in-time estimates when using modulated free energy instead of relative entropy is the modulated log-Sobolev inequalities discussed in [200]. These modulated LSI are in fact classical LSI satisfied uniformly over a specific family of measures (called the modulated Gibbs measures, and distinct from the law m_t that we consider in Theorem 4.19; but a similar time-uniformity is required). The arguments of the time-uniform LSI of Theorem 4.19 may thus be useful to establish time-uniform modulated LSI (although additional difficulties appear in the latter case, in particular a uniformity in the number of particles is required). On the topic of modulated free energy and modulated LSI, we mention that an upcoming work [119] is announced in [59].

Remark 4.21 (Non-conservative flow and more singularity). Two natural extensions to the setting considered in Assumption 4.14 are to consider a not necessarily anti-symmetric M (notably $M = -I_{d\times d}$ which corresponds to the gradient flow) and a more singular interaction with $s \in [d-1, d)$. We note that in the first case, a not anti-symmetric M poses challenges in mathematical analysis since the divergence term

$$\nabla \cdot (K \star m_t) = M : \nabla^2 g \star m_t$$

appears and is more singular than the flow m_t itself when $s \geqslant d-2$. By re-examining the proof, we find that the method of Theorem 4.19 will continue to work if $(m_t)_{t\geqslant 0}$ satisfies a uniform θ-Hölder bound for some $\theta > s-d+3$ without the anti-symmetry of M, or for some $\theta > s-d+2$ with an anti-symmetric M. The authors are unfortunately unaware of such results for Riesz flows with confinement in the whole space, which are possibly worthy of independent studies in the future.

4.4.3 Global PoC for log interaction with general confinement potential

As a consequence of Theorem 4.19, we get the strong uniform-in-time propagation of chaos result.

Theorem 4.22. *Let Assumption 4.14 hold and suppose additionally that $s = 0$ and M is anti-symmetric. Let $(m_t)_{t\geqslant 0}$ be a solution to (4.38) whose initial value m_0 satisfies the conditions of Theorem 4.19 and let $(m^N_t)_{t\geqslant 0}$ be a solution to (4.40). Then, there exist C, $\rho > 0$, depending only on d, U, $|M|$ and m_0, such that*

$$\mathcal{H}\big(m^N_t \big| m_t^{\otimes N}\big) \leqslant C \exp\big(-(\rho - C|M|)t\big)\mathcal{H}\big(m^N_0 \big| m_0^{\otimes N}\big) + C\Big(1 + \exp\big(-(\rho - C|M|)t\big)\Big) \tag{4.47}$$

for all $t \geqslant 0$, once $\mathcal{H}\big(m^N_0 \big| m_0^{\otimes N}\big) < \infty$. Moreover, when $|M|$ increases and all other dependencies are kept constant, C increases and ρ decreases.

By the monotonicity of C and ρ in $|M|$, we find $C|M| < \rho$ when $|M|$ is sufficiently small, and in this case the bound (4.47) becomes uniform in time. Even when $|M|$ is not small, we get a global PoC estimate for the dissipative log-interaction on $\mathbb{R}^d$ with a confinement potential, which is new to our knowledge (the case $U = 0$ is addressed in [90]).

The proof of Theorem 4.22 is postponed to Section 4.4.5.

4.4.4 Uniform PoC for log interaction with quadratic confinement potential

In this subsection, we impose the additional assumption.

Assumption 4.23. The confinement potential reads $U(x) = \kappa_U |x|^2/2$ for some $\kappa_U > 0$.

Under Assumptions 4.14 and 4.23, we easily verify that the Gaussian measure m_* with density

$$m_*(x) = \exp\bigl(-U(x)\bigr) = \exp\biggl(-\frac{\kappa_U |x|^2}{2}\biggr)$$

is invariant to the mean field flow (4.38). The first result that we obtain is the exponential convergence of the mean field flow towards m_*.

Theorem 4.24. *Let Assumptions 4.14 and 4.23 hold and suppose additionally that M is anti-symmetric. Let $m_0 \in \mathcal{P}(\mathbb{R}^d)$ satisfy the conditions of Theorem 4.19 and let $(m_t)_{t\geqslant 0}$ be the solution to (4.38). Then, we have, for all $t \geqslant 0$,*

$$\mathcal{H}(m_t|m_*) \leqslant \exp(-2\kappa_U t)\mathcal{H}(m_0|m_*)\,.$$

Moreover, setting $u_t = -\ln \mathrm{d}m_t/\mathrm{d}m_*$, we have, for all $t > 0$,

$$\sup_{x\in\mathbb{R}^d} |x \cdot K \star (m_t - m_*)| \leqslant Ce^{-ct}\,, \quad \|\nabla u\|_{L^\infty} \leqslant Ce^{-ct}\,, \quad \|\nabla^2 u\|_{L^\infty} \leqslant \frac{Ce^{-ct}}{\sqrt{t \wedge 1}}$$

for some C, $c > 0$ that depend only on d, s, κ_U, M and $\|\nabla u_0\|_{L^\infty}$.

The proof of Theorem 4.24 is postponed to Section 4.4.5.

Building upon the exponential convergence above, we obtain the uniform-in-time propagation of chaos without restriction on the strength of the interaction.

Theorem 4.25. *Let Assumptions 4.14 and 4.23 hold and suppose additionally that $s = 0$ and M is anti-symmetric. Let $(m_t)_{t\geqslant 0}$ be a solution to (4.38) whose initial value m_0 satisfies the conditions of Theorem 4.19 and let $(m_t^N)_{t\geqslant 0}$ be a solution to (4.40). Then, there exists $C > 0$, depending only on d, κ_U, M and m_0, such that*

$$\mathcal{H}\bigl(m_t^N\big|m_t^{\otimes N}\bigr) \leqslant C\exp(-2\kappa_U t)\Bigl(\mathcal{H}\bigl(m_0^N\big|m_0^{\otimes N}\bigr) + 1\Bigr) \tag{4.48}$$

for all $t \geqslant 0$, once $\mathcal{H}\bigl(m_0^N\big|m_0^{\otimes N}\bigr) < \infty$.

The proof of Theorem 4.25 is postponed to Section 4.4.5. Notice that, as discussed in e.g. [200], this result describes a *generation* of chaos property (not only *propagation*) since it implies that $\mathcal{H}\bigl(m_t^N\big|m_t^{\otimes N}\bigr)$ is of order 1 (in terms of N) for large times even if it is not the case at time $t = 0$. Here, moreover, and more surprisingly, the right hand side of (4.48) vanishes at $t \to \infty$, which is due to the fact that in the specific case of an isotropic Gaussian confining potential, the invariant measure of the system of interacting particles is a tensorized Gaussian distribution, which is thus also the long-time limit of the product of solutions of the non-linear equation. Finally, in contrast with the results stated in Section 4.3, here (as in Theorem 4.22) we only state a result on the relative entropy of the full system, and thus by sub-additivity of the relative entropy this yields PoC estimates on the k-particles marginals which are not sharp in the sense of [140, 142].

4.4.5 Proofs

Proofs of uniform bounds and LSI

Proof of Theorem 4.19. Set $\mu_0 = Z^{-1}\exp(-U)$ with $Z = \int \exp(-U)$.

Step 1: Construction of a regular approximation. Recall that the initial condition m_0 is such that

$$u_0 = -\ln \frac{\mathrm{d}m_0}{\mathrm{d}\mu_0} = -\ln m_0 - \ln Z - U$$

is Lipschitz continuous. We construct, for $\varepsilon > 0$, the approximative initial value

$$m_0^\varepsilon = \frac{\exp(-u_0 \star \eta^\varepsilon)\mu_0}{\int \exp(-u_0 \star \eta^\varepsilon)\mu_0^\varepsilon},$$

where $\mu_0^\varepsilon \propto \exp(-U^\varepsilon)$. Construct as well the approximative dynamics (4.42) with the mollified kernel $K^\varepsilon = K \star \eta^\varepsilon$ and mollified confinement $U^\varepsilon = U \star \eta^\varepsilon$. By construction, the initial value m_0^ε converges to m_0 in L^1 and is uniformly bounded in L^∞, thus the last claim of Proposition 4.16 indicates that $m_t^\varepsilon \to m_t$ in L^1 for all $t \geqslant 0$. Using the uniqueness of the solution of the Fokker–Planck equation satisfied by the relative density $\mathrm{d}m_t^\varepsilon/\mathrm{d}\mu_0^\varepsilon$ and a Feynman–Kac argument similar to that of Proposition C.2, we obtain that $u_t^\varepsilon \coloneqq -\ln \mathrm{d}m_t^\varepsilon/\mathrm{d}\mu_0^\varepsilon$ is $\mathcal{C}^2$ in space. As a consequence, u_t^ε is a classical solution to the HJB equation

$$\partial_t u_t^\varepsilon = \Delta u_t^\varepsilon - |\nabla u_t^\varepsilon|^2 + \tilde{b}_t^\varepsilon \cdot \nabla u_t^\varepsilon + \varphi_t^\varepsilon,$$

where $\tilde{b}_t^\varepsilon$, φ_t^ε are given by

$$\begin{aligned}\tilde{b}_t^\varepsilon &= -\nabla U^\varepsilon - K^\varepsilon \star m_t^\varepsilon, \\ \varphi_t^\varepsilon &= -\nabla \cdot (K^\varepsilon \star m_t^\varepsilon) - (K^\varepsilon \star m_t^\varepsilon)\cdot \nabla U^\varepsilon.\end{aligned}$$

Let $u_0'^\varepsilon = u_0^\varepsilon$, and let $u_t'^\varepsilon$ denote the unique classical solution to this equation as established by Theorem 4.18. By considering the Fokker–Planck equation satisfied by $\mu_0^\varepsilon \exp(-u_t'^\varepsilon)$ and invoking the uniqueness of its solution, it follows that $u_t'^\varepsilon = u_t^\varepsilon$. Consequently, the regularity bounds stated in Theorem 4.18 are applicable to u_t^ε.

Step 2: Uniform bound on $K\star m_t$ and ∇u_t, and uniform LSI. We verify that the drift $\tilde{b}_t^\varepsilon$ satisfies the semi-monotonicity condition of Theorem 4.18, as the contribution from the interaction $K^\varepsilon \star m_t^\varepsilon$ is controlled by Proposition 4.15:

$$\|K^\varepsilon \star m_t^\varepsilon\|_{L^\infty} \leqslant \|K \star m_t^\varepsilon\|_{L^\infty} \lesssim \|m_t^\varepsilon\|_{L^1}^{1-(s+1)/d}\|m_t^\varepsilon\|_{L^\infty}^{(s+1)/d},$$

and U (along with its approximation U^ε) is already weakly convex. Now we focus on proving the uniform L^∞ bound on φ_t^ε. For the first term in φ_t^ε, we find that in the Coulombic and the super-Coulombic cases, due to the anti-symmetry of M, we have

$$\nabla \cdot (K^\varepsilon \star m_t^\varepsilon) = \nabla \cdot (M \nabla g \star m_t^\varepsilon \star \eta^\varepsilon) = M : g \star \nabla^2 (m_t^\varepsilon \star \eta^\varepsilon) = 0\,;$$

for the sub-Coulombic case where $s < d-2$, applying Proposition 4.15 with $p = \infty$, we get

$$\|\nabla \cdot (K^\varepsilon \star m_t^\varepsilon)\|_{L^\infty} \lesssim \|m_t^\varepsilon\|_{L^1}^{1-(s+2)/d} \|m_t^\varepsilon\|_{L^\infty}^{(s+2)/d}\,,$$

so the first term is uniformly bounded in L^∞ in both cases. To treat the second term, we note that

$$|K^\varepsilon \star m^\varepsilon(x)| \leqslant \sup_{x' \in B(x,\varepsilon)} |K \star m^\varepsilon(x')|\,,$$

so it suffices to prove the bound uniformly:

$$|K \star m^\varepsilon(x')| \lesssim (1 + |x|)^{-1}\,.$$

Decompose the kernel in the following way:

$$K(x) = K(x)\mathbb{1}_{|x|<R} + K(x)\mathbb{1}_{|x|\geqslant R} =: K_1(x) + K_2(x)\,.$$

For the exploding part K_1, we have

$$\begin{aligned}
|K_1 \star m_t^\varepsilon(x)| &= \left| \int_{B(x,R)} K(x-y) m_t^\varepsilon(y)\,\mathrm{d}y \right| \\
&\lesssim \int_{B(x,R)} |x-y|^{-s-1} m_t^\varepsilon(y)\,\mathrm{d}y \\
&\leqslant \left(\int_{B(x,R)} |x-y|^{-p(s+1)}\,\mathrm{d}y \right)^{1/p} \|m_t^\varepsilon \mathbb{1}_{B(x,R)}\|_{L^q} \\
&\lesssim R^{d/p-s-1} \|m_t^\varepsilon \mathbb{1}_{B(x,R)}\|_{L^q}\,,
\end{aligned}$$

where $p \in \left(1, \frac{d}{s+1}\right)$ and $p^{-1} + q^{-1} = 1$. For $|x| > R$, we observe

$$\int_{B(x,R)} (m_t^\varepsilon)^q \leqslant \|m_t^\varepsilon\|_{L^\infty}^{q-1} \int_{B(x,R)} m_t^\varepsilon \leqslant \|m_t^\varepsilon\|_{L^\infty}^{q-1} (|x| - R)^{-q} \int_{\mathbb{R}^d} |x|^q m_t^\varepsilon(x)\,\mathrm{d}x\,.$$

For the non-exploding part K_2, we have

$$\begin{aligned}
|K_2 \star m_t^\varepsilon(x)| &= \left| \int_{\mathbb{R}^d \setminus B(x,R)} K(x-y) m_t^\varepsilon(y)\,\mathrm{d}y \right| \\
&\lesssim \int_{\mathbb{R}^d \setminus B(x,R)} |x-y|^{-s-1} m_t^\varepsilon(y)\,\mathrm{d}y \\
&= |x|^{-s-1} \int_{\mathbb{R}^d \setminus B(x,R)} \frac{|x|^{s+1}}{|x-y|^{s+1}} m_t^\varepsilon(y)\,\mathrm{d}y \\
&\lesssim |x|^{-s-1} \int_{\mathbb{R}^d \setminus B(x,R)} \frac{|x-y|^{s+1} + |y|^{s+1}}{|x-y|^{s+1}} m_t^\varepsilon(y)\,\mathrm{d}y \\
&\leqslant |x|^{-s-1} \int_{\mathbb{R}^d} (1 + R^{-s-1}|y|^{s+1}) m_t^\varepsilon(y)\,\mathrm{d}y\,.
\end{aligned}$$

Thanks to Proposition 4.16, the mean field flow $(m_t^\varepsilon)_{t\geqslant 0}$ enjoys uniform bounds on its L^∞ norm and its moments, as all moments of its initial value m_0^ε are finite. Thus, we have, uniformly in t,

$$\sup_{t\geqslant 0}|K\star m_t^\varepsilon(x)| \leqslant \sup_{t\geqslant 0}|K_1\star m_t^\varepsilon(x) + K_2\star m_t^\varepsilon(x)| \lesssim (1+|x|)^{-1}.$$

So we have obtained $\sup_{t\geqslant 0}\|\varphi_t^\varepsilon\|_{L^\infty} < \infty$, and the first claim of Theorem 4.18 implies that $\|\nabla u_t^\varepsilon\|_{L^\infty}$ is uniformly bounded. Taking the limit $\varepsilon\to 0$, we recover the uniform spatial Lipschitz bound on u_t and thus by the perturbation lemma of Aida–Shigekawa, the flow $(m_t)_{t\geqslant 0}$ satisfies a uniform LSI.

Step 3: Uniform bound on $\nabla^2 u_t$. We want to apply the second claim of Theorem 4.18 to the HJB solution u_t^ε, and it suffices to control uniformly in time the following quantities:

$$\begin{aligned}\nabla\tilde{b}_t^\varepsilon\cdot\nabla u_t^\varepsilon &= (-\nabla^2 U^\varepsilon - K^\varepsilon\star\nabla m_t^\varepsilon)\cdot\nabla u_t^\varepsilon,\\ \nabla\varphi_t^\varepsilon &= -\nabla^2\cdot(K^\varepsilon\star m_t^\varepsilon) - \nabla(K^\varepsilon\star m_t^\varepsilon)\cdot\nabla U^\varepsilon - (K^\varepsilon\star m_t^\varepsilon)\cdot\nabla^2 U^\varepsilon.\end{aligned}$$

The first quantity can be bounded by

$$\|\nabla\tilde{b}_t^\varepsilon\cdot\nabla u_t^\varepsilon\|_{L^\infty} \leqslant \|\nabla\tilde{b}_t^\varepsilon\|_{L^\infty}\|\nabla u_t^\varepsilon\|_{L^\infty} \leqslant \bigl(\|\nabla^2 U\|_{L^\infty} + \|K^\varepsilon\star\nabla m_t^\varepsilon\|_{L^\infty}\bigr)\|\nabla u_t^\varepsilon\|_{L^\infty},$$

where $\|K^\varepsilon\star\nabla m_t\|_{L^\infty}$ is uniformly bounded as

$$\nabla m_t^\varepsilon = m_t^\varepsilon(-\nabla U^\varepsilon + \nabla u_t^\varepsilon) = \frac{\exp(-U^\varepsilon - u_t^\varepsilon)}{\int\exp(-U^\varepsilon - u_t^\varepsilon)}(-\nabla U^\varepsilon + \nabla u_t^\varepsilon) \in L^1\cap L^\infty$$

uniformly in time, thanks to the uniform bound on ∇u_t^ε. Now consider the second quantity $\nabla\varphi_t^\varepsilon$. In the case $s\in[d-2,d-1)$, we have $K = M\nabla g$ with an anti-symmetric M, so the first term $\nabla^2\cdot(K^\varepsilon\star m_t^\varepsilon)$ vanishes. In the case $s<d-2$, we have

$$\|\nabla^2\cdot(K^\varepsilon\star m_t^\varepsilon)\|_{L^\infty} \leqslant \|\nabla K\star\nabla m_t^\varepsilon\|_{L^\infty} \lesssim \|\nabla m_t^\varepsilon\|_{L^1}^{1-(s+2)/d}\|\nabla m_t^\varepsilon\|_{L^\infty}^{(s+2)/d},$$

and by the uniform L^1 and L^∞ bound on ∇m_t^ε, this term is uniformly bounded. That is to say, in both cases, the first term $\nabla^2\cdot(K^\varepsilon\star m_t^\varepsilon)$ is uniformly bounded in L^∞. As we have $\|\nabla^2 U\|_{L^\infty}<\infty$, the third term $(K^\varepsilon\star m_t^\varepsilon)\cdot\nabla^2 U$ is equally uniformly bounded. So it remains to obtain a uniform bound on the second term $\nabla(K^\varepsilon\star m_t^\varepsilon)\cdot\nabla U$. Since ∇U^ε is of linear growth, it suffices to prove

$$\nabla(K\star m_t^\varepsilon)(x) = (K\star\nabla m_t^\varepsilon)(x) \lesssim (1+|x|)^{-1}$$

uniformly in time. For this, we use again the decomposition $K = K_1 + K_2$ in the end of the previous step, and redoing all the computations, we find that it is sufficient to uniformly control

$$\begin{aligned}\int_{\mathbb{R}^d}|x|^q|\nabla m_t^\varepsilon(x)|\,\mathrm{d}x &= \int_{\mathbb{R}^d}|x|^q|-\nabla U(x) + \nabla u_t^\varepsilon(x)|m_t^\varepsilon(x)\,\mathrm{d}x\\ &\lesssim \int_{\mathbb{R}^d}|x|^q(1+|x|)m_t^\varepsilon(x)\,\mathrm{d}x\end{aligned}$$

for some $q > \bigl(1-\frac{s+1}{d}\bigr)^{-1}$. But from Proposition 4.16 we know that the q and $(q+1)$-th moments of m_t^ε are uniformly bounded. Hence, $\nabla\varphi_t^\varepsilon$ is uniformly bounded in L^∞ and by the second claim of Theorem 4.18, we get that $\|\nabla^2 u_t^\varepsilon\|_{L^\infty}$ is uniformly bounded. Thus $\nabla^2\ln m_t^\varepsilon = -\nabla^2 U^\varepsilon - \nabla^2 u_t^\varepsilon$ is uniformly bounded as well, and taking the limit $\varepsilon\to 0$, we get the desired result for u_t. □

Proof of Theorem 4.24. The proof is similar to that of Theorem 4.19, except that now the Lipschitz and Hessian bounds converge to zero. Thus, we first show that the mean field flow m_t converges to the invariant measure m_* and then redo the estimates on the log-density.

Step 1: Convergence in entropy. For the initial value m_0 such that

$$\nabla u_0 = -\nabla \ln m_0 - \nabla U \in L^\infty,$$

we find, as in the beginning of the proof of Theorem 4.19, an approximation m_0^ε defined by the following:

$$m_0^\varepsilon = \frac{\exp(-u_0 \star \eta^\varepsilon - U)}{\int \exp(-u_0 \star \eta^\varepsilon - U)}\,.$$

Set $u_t^\varepsilon \coloneqq -\ln \mathrm{d}m_0^\varepsilon/\mathrm{d}m_*$. We also consider the approximative flow $(m_t^\varepsilon)_{t\geqslant 0}$ solving the mean field Fokker–Planck equation (4.42). Notice that, in the case of quadratic potential, the mollified potential $U^\varepsilon = U \star \eta^\varepsilon$ is nothing but U translated by a constant, due to the symmetry of η^ε. By Feynman–Kac arguments, we get that m_t^ε is a classical solution to the Fokker–Planck and $\nabla^i u_t^\varepsilon$ grows at most linearly for $i = 0$, 1, 2. Thus, we can derive $t \mapsto \mathcal{H}(m_t^\varepsilon|m_*)$ and get

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{H}(m_t^\varepsilon|m_*)}{\mathrm{d}t} &= -\mathcal{I}(m_t^\varepsilon|m_*) + \int_{\mathbb{R}^d} \nabla \ln \frac{m_t^\varepsilon(x)}{m_*(x)} \cdot K^\varepsilon \star (m_t^\varepsilon - m_*)(x) m_t^\varepsilon(\mathrm{d}x) \\
&= -\mathcal{I}(m_t^\varepsilon|m_*) - \int_{\mathbb{R}^d} \nabla \ln m_*(x) \cdot K^\varepsilon \star (m_t^\varepsilon - m_*)(x) m_t^\varepsilon(\mathrm{d}x) \\
&= -\mathcal{I}(m_t^\varepsilon|m_*) - \int_{\mathbb{R}^d} \nabla \ln m_*(x) \cdot K^\varepsilon \star m_t^\varepsilon(x) m_t^\varepsilon(\mathrm{d}x) \\
&= -\mathcal{I}(m_t^\varepsilon|m_*) + \kappa_U \int x \cdot K^\varepsilon \star m_t^\varepsilon(x) m_t(\mathrm{d}x) \\
&= -\mathcal{I}(m_t^\varepsilon|m_*) + \kappa_U \iint_{\mathbb{R}^d\times\mathbb{R}^d} x \cdot K^\varepsilon(x-y) m_t^\varepsilon(\mathrm{d}x) m_t^\varepsilon(\mathrm{d}y) \\
&= -\mathcal{I}(m_t^\varepsilon|m_*) + \frac{\kappa_U}{2} \iint_{\mathbb{R}^d\times\mathbb{R}^d} (x-y) \cdot K^\varepsilon(x-y) m_t^\varepsilon(\mathrm{d}x) m_t^\varepsilon(\mathrm{d}y) \\
&= -\mathcal{I}(m_t^\varepsilon|m_*) \leqslant -2\kappa_U \mathcal{H}(m_t^\varepsilon|m_*)\,.
\end{aligned}$$

Here the second inequality is due to the integration by parts and the fact that $\nabla \cdot K^\varepsilon = 0$; the third to the fact that $\nabla \ln m_*(x)$ is parallel to x and $K^\varepsilon \star m_*(x) = K \star (m_* \star \eta^\varepsilon)(x)$ is always orthogonal to x, as $m_* \star \eta^\varepsilon$ is invariant by rotation; the sixth due to the oddness of K^ε; and the last due to

$$x \cdot K^\varepsilon(x) = x^\top M \nabla g^\varepsilon(x)$$

and $\nabla g^\varepsilon(x)$ is parallel to x. Then applying Grönwall's lemma and the log-Sobolev inequality for m_*, we get

$$\mathcal{H}(m_t^\varepsilon|m_*) \leqslant \mathcal{H}(m_0^\varepsilon|m_*) \exp(-2\kappa_U t)\,,$$

and taking the limit $\varepsilon \to 0$ and using the lower semi-continuity of relative entropy, we recover the first claim.

Step 2: Decaying bound on $x \cdot K \star m_t(x)$ and ∇u_t. In the following, C, c will denote positive reals that has the same dependency as stated in the theorem and may change from line to line. Working again with the approximation m_t^ε, we get by Pinsker's inequality,

$$\begin{aligned}\|m_t^\varepsilon - m_*\|_{L^1} &\leqslant \exp(-\kappa_U t)\sqrt{2\mathcal{H}(m_0^\varepsilon|m_*)} \\ &\leqslant \exp(-\kappa_U t)\sqrt{2\kappa_U^{-1}\mathcal{I}(m_0^\varepsilon|m_*)} \\ &\leqslant \exp(-\kappa_U t)\sqrt{2\kappa_U^{-1}\|\nabla u_0^\varepsilon\|_{L^\infty}^2} = C\exp(-\kappa_U t)\,.\end{aligned}$$

Then, applying Proposition 4.15, we get

$$\|K^\varepsilon \star (m_t^\varepsilon - m_*)\|_{L^\infty} \leqslant C\|m_t^\varepsilon - m_*\|_{L^1}^{1-(s+1)/d}\|m_t^\varepsilon - m_*\|_{L^\infty}^{(s+1)/d} \leqslant Ce^{-ct}\,.$$

We know that $u_t^\varepsilon = -\ln \mathrm{d}m_t^\varepsilon/\mathrm{d}m_*$ solves the HJB equation

$$\partial_t u_t^\varepsilon = \Delta u_t^\varepsilon - |\nabla u_t^\varepsilon|^2 + \tilde{b}_t^\varepsilon \cdot \nabla u_t + \varphi_t^\varepsilon$$

for $\tilde{b}_t^\varepsilon(x) = -\kappa_U x - K^\varepsilon \star m_t^\varepsilon(x)$ and $\varphi_t^\varepsilon(x) = -\kappa_U x \cdot K^\varepsilon \star m_t^\varepsilon(x)$. Note that φ_t satisfies

$$\varphi_t^\varepsilon(x) = -\kappa_U x \cdot K^\varepsilon \star (m_t^\varepsilon - m_*)(x)\,,$$

since $x \cdot K^\varepsilon \star m_*(x) = 0$ according to the argument in Step 1. Thus, we have

$$\begin{aligned}\varphi_t^\varepsilon(x) &= -\kappa_U \int_{\mathbb{R}^d} x^\top M\nabla g^\varepsilon(x-y)(m_t^\varepsilon - m_*)(\mathrm{d}y) \\ &= -\kappa_U \int_{\mathbb{R}^d} y^\top M\nabla g^\varepsilon(x-y)(m_t^\varepsilon - m_*)(\mathrm{d}y)\,,\end{aligned}$$

as $x^\top M\nabla g^\varepsilon(x) = 0$ for all $x \in \mathbb{R}^d$. So φ_t^ε satisfies the bound

$$\begin{aligned}\|\varphi_t^\varepsilon\|_{L^\infty} &= \left|\int_{\mathbb{R}^d} y^\top K^\varepsilon(x-y)(m_t^\varepsilon - m_*)(\mathrm{d}y)\right| \\ &\lesssim \int_{B(0,1)} \frac{|y|}{|x-y|^{s+1}}|m_t^\varepsilon - m_*|(\mathrm{d}y) + \sup_{y:|y-x|\geqslant 1}|y^\top K^\varepsilon(x-y)|\,\|m_t^\varepsilon - m_*\|_{L^1} \\ &\lesssim \|(m_t^\varepsilon - m_*)\mathbb{1}_{B(x,1)}\|_{L^q} + \|m_t^\varepsilon - m_*\|_{L^1}\end{aligned}$$

for $q > \left(1 - \frac{s+1}{d}\right)^{-1}$, according to the argument in the proof of Theorem 4.19. For the L^q norm we have, by interpolation,

$$\|(m_t^\varepsilon - m_*)\mathbb{1}_{B(x,1)}\|_{L^q} \leqslant \|m_t^\varepsilon - m_*\|_{L^q} \leqslant \|m_t^\varepsilon - m_*\|_{L^1}^{1/q}\|m_t^\varepsilon - m_*\|_{L^\infty}^{1/p} \lesssim \|m_t^\varepsilon - m_*\|_{L^1}^{1/q}$$

for $p^{-1} + q^{-1} = 1$. Thus, $\|\varphi_t^\varepsilon\|_{L^\infty} \leqslant Ce^{-ct}$, and applying the first claim of Theorem 4.18, we get $\|\nabla u_t^\varepsilon\|_{L^\infty} \leqslant Ce^{-ct}$. The first claim is then proved by taking the limit $\varepsilon \to 0$.

Step 3: Decaying bound on $\nabla^2 u_t$. First, we have

$$\|\nabla\tilde{b}_t^\varepsilon \cdot \nabla u_t^\varepsilon\|_{L^\infty} \leqslant \|\nabla\tilde{b}_t^\varepsilon\|_{L^\infty} \cdot Ce^{-ct} \leqslant \left(\|\nabla^2 U\|_{L^\infty} + \|K^\varepsilon \star \nabla m_t^\varepsilon\|_{L^\infty}\right) \cdot Ce^{-ct}\,,$$

where

$$\|K^\varepsilon \star \nabla m_t^\varepsilon\|_{L^\infty} \lesssim \|\nabla m_t^\varepsilon\|_{L^1}^{1-(s+1)/d}\|\nabla m_t^\varepsilon\|_{L^\infty}^{(s+1)/d}\,.$$

As we have

$$\nabla m_t^\varepsilon = -(\nabla U + \nabla u_t^\varepsilon) m_t^\varepsilon = -\frac{\nabla U \exp(-U - u_t^\varepsilon)}{\int \exp(-U - u_t^\varepsilon)} - \nabla u_t^\varepsilon \, m_t^\varepsilon$$

with ∇U of linear growth and ∇u_t^ε being uniformly bounded, we find that $\nabla m_t^\varepsilon \in L^1 \cap L^\infty$ uniformly. Thus,

$$\|\nabla \tilde{b}_t^\varepsilon \cdot \nabla u_t^\varepsilon\|_{L^\infty} \leqslant Ce^{-ct} .$$

The gradient of φ_t^ε reads

$$\begin{aligned} \nabla \varphi_t^\varepsilon(x) &= -\nabla\big(\kappa_U x \cdot K^\varepsilon \star (m_t^\varepsilon - m_*)(x)\big) \\ &= -\kappa_U K^\varepsilon \star (m_t^\varepsilon - m_*)(x) - \kappa_U x \cdot K^\varepsilon \star \nabla (m_t^\varepsilon - m_*)(x) . \end{aligned}$$

The first term on the right hand side is already controlled:

$$|K^\varepsilon \star (m_t - m_*)(x)| \lesssim \|m_t - m_*\|_{L^1}^{1-(s+1)/d} \|m_t - m_*\|_{L^\infty}^{(s+1)/d} \leqslant Ce^{-ct} ,$$

and in the following we show that the same is true for the second term. Again, using the fact that $x \cdot K^\varepsilon(x) = 0$, we get

$$\begin{aligned} x \cdot K^\varepsilon \star \nabla (m_t^\varepsilon - m_*)(x) &= \int_{\mathbb{R}^d} x^\top M \nabla g^\varepsilon(x-y) \nabla (m_t^\varepsilon - m_*)(\mathrm{d}y) \\ &= \int_{\mathbb{R}^d} y^\top M \nabla g^\varepsilon(x-y) \nabla (m_t^\varepsilon - m_*)(\mathrm{d}y) . \end{aligned}$$

Following the argument in Step 2, we separate the two cases $|y-x| < 1$ and $\geqslant 1$, and get

$$\sup_{x \in \mathbb{R}^d} |x \cdot K^\varepsilon \star \nabla (m_t^\varepsilon - m_*)(x)| \lesssim \|\nabla (m_t^\varepsilon - m_*)\|_{L^q} + \|\nabla (m_t^\varepsilon - m_*)\|_{L^1}$$

for $q > \big(1 - \frac{s+1}{d}\big)^{-1}$. Using the explicit density of m_t^ε, we get

$$\nabla m_t^\varepsilon - \nabla m_* = -\nabla u_t^\varepsilon \, m_t^\varepsilon - \nabla U (m_t^\varepsilon - m_*) .$$

The first term satisfies

$$\|\nabla u_t^\varepsilon \, m_t^\varepsilon\|_{L^1} \leqslant \|\nabla u_t^\varepsilon\|_{L^\infty} \|m_t^\varepsilon\|_{L^1} \leqslant Ce^{-ct} ,$$

and the second satisfies

$$\|\nabla U (m_t^\varepsilon - m_*)\|_{L^1} \leqslant \|\nabla^2 U\|_{L^\infty} W_1(m_t^\varepsilon, m_*) \lesssim \sqrt{\mathcal{H}(m_t^\varepsilon | m_*)} \leqslant Ce^{-ct} .$$

Finally, their densities have the L^∞ bounds:

$$\begin{aligned} \|\nabla u_t^\varepsilon \, m_t^\varepsilon\|_{L^\infty} &\leqslant \|\nabla u_t^\varepsilon\|_{L^\infty} \|m_t^\varepsilon\|_{L^\infty} \leqslant C , \\ \|\nabla U (m_t^\varepsilon - m_*)\|_{L^\infty} &\lesssim \sup_{x \in \mathbb{R}^d} (1 + |x|) \left(\frac{\exp\big(-U(x)\big)}{\int \exp(-U)} + \frac{\exp\big(-u_t^\varepsilon(x) - U(x)\big)}{\int \exp(-u_t^\varepsilon - U)} \right) \leqslant C . \end{aligned}$$

Then, by the same interpolation as in Step 2, we get $\|\nabla \varphi_t^\varepsilon\|_{L^\infty} \leqslant Ce^{-ct}$. Applying the second claim of Theorem 4.18, we get

$$\|\nabla^2 u_t^\varepsilon\|_{L^\infty} \leqslant \frac{Ce^{-ct}}{\sqrt{t \wedge 1}} + \int_0^t \frac{Ce^{-cv}}{\sqrt{v \wedge 1}} \cdot Ce^{-c(t-v)} \, \mathrm{d}v \leqslant \frac{Ce^{-ct}}{\sqrt{t \wedge 1}} .$$

Taking the limit $\varepsilon \to 0$, we recover the second claim and this concludes the proof. $\square$

Proofs of propagation of chaos

Proof of Theorem 4.22. According to Propositions 4.16 and 4.17, given the initial values m_0, m_0^N, we find respectively approximating sequences m_0^ε, $m_0^{\varepsilon,N}$ such that $\ln m_0^\varepsilon + U^\varepsilon \in \mathcal{C}_\mathrm{b}^\infty$ and $\boldsymbol{x} \mapsto \ln m_0^{\varepsilon,N}(\boldsymbol{x}) + \sum_{i=1}^N U^\varepsilon(x^i) \in \mathcal{C}_\mathrm{b}^\infty$. The solutions of (4.42) and of the forward Kolmogorov equation associated to (4.43) being unique, we can use the Feynman–Kac representation of Proposition C.2 to find that the densities and their classically derivatives

$$\nabla^i\bigl(\ln m_t^\varepsilon + U^\varepsilon\bigr)\,,\quad \nabla^i\Biggl(\ln m_t^{\varepsilon,N}(\boldsymbol{x}) + \sum_{i=1}^N U^\varepsilon(x^i)\Biggr)\,,\qquad i \geqslant 1$$

exist and grow at most linearly in space (locally in time). Then in the following we can justify all the exchanges between limit and integration, and all the integrations by parts. Taking the derivative of the relative entropy $\mathcal{H}_t^\varepsilon = \mathcal{H}\bigl(m_t^{\varepsilon,N}\big|(m_t^\varepsilon)^{\otimes N}\bigr)$, and denoting the relative Fisher information by

$$\mathcal{I}_t^\varepsilon = \mathcal{I}\bigl(m_t^{\varepsilon,N}\big|(m_t^\varepsilon)^{\otimes N}\bigr) = \int_{\mathbb{R}^{dN}} \left|\nabla \ln \frac{m_t^{\varepsilon,N}(\boldsymbol{x})}{(m_t^\varepsilon)^{\otimes N}(\boldsymbol{x})}\right|^2 m_t^{\varepsilon,N}(\mathrm{d}\boldsymbol{x})\,,$$

we get

$$\begin{aligned}\frac{\mathrm{d}\mathcal{H}_t^\varepsilon}{\mathrm{d}t} &= -\mathcal{I}_t^\varepsilon + \frac{1}{N-1}\sum_{i\neq j}\int_{\mathbb{R}^{dN}} \nabla_i \ln\frac{m_t^{\varepsilon,N}(\boldsymbol{x})}{m_t^\varepsilon(x^i)} \\ &\qquad\qquad\qquad\qquad \cdot \bigl(K^\varepsilon(x^i - x^j) - K^\varepsilon \star m_t^\varepsilon(x^i)\bigr) m_t^{\varepsilon,N}(\mathrm{d}\boldsymbol{x}) \\ &= -\mathcal{I}_t^\varepsilon - \frac{1}{N-1}\sum_{i\neq j}\int_{\mathbb{R}^{dN}} \nabla \ln m_t^\varepsilon(x^i) \\ &\qquad\qquad\qquad\qquad \cdot \bigl(K^\varepsilon(x^i - x^j) - K^\varepsilon \star m_t^\varepsilon(x^i)\bigr) m_t^{\varepsilon,N}(\mathrm{d}\boldsymbol{x})\,,\end{aligned}$$

where i, j are summed over $[\![1,N]\!]$ and the second equality is due to integration by parts and the fact that $\nabla\cdot K^\varepsilon = 0$. Noting that the regularized N-particle measure $m_t^{\varepsilon,N}$ has density and has no mass on the sets $\{\boldsymbol{x} : x^i = x^j\}$ for $i \neq j$, we find that the second term is equal to, by symmetrization,

$$-\frac{1}{N-1}\sum_{i,j=1}^N \int_{\mathbb{R}^{dN}} \phi_t(x^i, x^j) m_t^{\varepsilon,N}(\mathrm{d}\boldsymbol{x})\,,$$

where the function $\phi_t(\cdot,\cdot)$ is given by

$$\begin{aligned}\phi_t(x,y) &= \frac{1}{2}K^\varepsilon(x-y)\cdot\bigl(\nabla\ln m_t^\varepsilon(x) - \nabla \ln m_t^\varepsilon(y)\bigr)\mathbb{1}_{x\neq y} \\ &\qquad - \frac{1}{2}K^\varepsilon\star m_t^\varepsilon(x)\cdot\nabla\ln m_t^\varepsilon(x) - \frac{1}{2}K^\varepsilon \star m_t^\varepsilon(y)\cdot\nabla\ln m_t^\varepsilon(y)\,.\end{aligned} \tag{4.49}$$

The function ϕ_t satisfies

$$\begin{aligned}\int_{\mathbb{R}^d} \phi_t(x,y) m_t^\varepsilon(\mathrm{d}y) &= 0\,,\\ \int_{\mathbb{R}^d} \phi_t(y,x) m_t^\varepsilon(\mathrm{d}y) &= 0\end{aligned}$$

for all $x \in \mathbb{R}^d$. From now on, the symbols C_i, $i \in \mathbb{N}$ denote a positive number that has the same dependency as C, ρ have in the statement of the theorem. For the first term in (4.49), we have by Theorem 4.19,

$$\sup_{x,y\,:\,x\neq y}\bigl|K^\varepsilon(x-y)\cdot\bigl(\nabla\ln m_t^\varepsilon(x)-\nabla\ln m_t^\varepsilon(y)\bigr)\bigr| \leqslant C_1|M|\|\nabla^2\ln m_t\|_{L^\infty} \leqslant \frac{C_2|M|}{\sqrt{t\wedge 1}}\,.$$

For the last two terms in the definition (4.49) of ϕ_t, we have by the same theorem,

$$|K\star m_t^\varepsilon(x)| \leqslant C_3|M|(1+|x|)^{-1}\,,$$
$$|\nabla\ln m_t^\varepsilon(x)| = |\nabla u_t^\varepsilon(x)| + |\nabla U(x)| \leqslant C_4(1+|x|)\,.$$

Thus,

$$|K\star m_t^\varepsilon(x)\cdot\nabla\ln m_t^\varepsilon(x)| \leqslant C_6|M|\,.$$

So the functions ϕ_t satisfies

$$\|\phi_t\|_{L^\infty} \leqslant \frac{C_7|M|}{\sqrt{t\wedge 1}}\,.$$

Therefore, using the convex duality for relative entropy, we get

$$\frac{\mathrm{d}\mathcal{H}_t^\varepsilon}{\mathrm{d}t} = -\mathcal{I}_t^\varepsilon + \delta_t\mathcal{H}_t^\varepsilon + \delta_t\ln\int_{\mathbb{R}^{dN}}\exp\biggl(\frac{1}{\delta_t(N-1)}\sum_{i,j=1}^N\phi_t(x^i,x^j)\biggr)(m_t^\varepsilon)^{\otimes N}(\mathrm{d}\boldsymbol{x})\,,$$

where we set

$$\delta_t = \frac{3(1600^2+36e^4)C_7|M|}{\sqrt{t\wedge 1}}\,.$$

Then, applying the "concentration" estimate [124, Theorem 4] (whose constant is given explicitly in [98, Theorem 5]), we obtain

$$\frac{\mathrm{d}\mathcal{H}_t^\varepsilon}{\mathrm{d}t} \leqslant -\mathcal{I}_t^\varepsilon + \frac{C_8|M|}{\sqrt{t\wedge 1}}\mathcal{H}_t^\varepsilon + \frac{C_8|M|}{\sqrt{t\wedge 1}} \leqslant -C_9\mathcal{H}_t^\varepsilon + \frac{C_8|M|}{\sqrt{t\wedge 1}}\mathcal{H}_t^\varepsilon + \frac{C_8|M|}{\sqrt{t\wedge 1}}\,.$$

We conclude by applying Grönwall's lemma and taking the limit $\varepsilon \to 0$. □

Proof of Theorem 4.25. The argument is largely the same as the proof above, i.e. the proof of Theorem 4.22. So here we only indicate the differences. Defining the same ϕ_t function as in (4.49), we find that in the quadratic case, we have the following bounds by Theorem 4.24:

$$\biggl\|\nabla\ln\frac{m_t^\varepsilon}{m_*}\biggr\|_{L^\infty} \leqslant C_1e^{-ct}\,,$$
$$\biggl\|\nabla^2\ln\frac{m_t^\varepsilon}{m_*}\biggr\|_{L^\infty} \leqslant \frac{C_1e^{-ct}}{\sqrt{t\wedge 1}}\,,$$
$$\sup_x|x\cdot K^\varepsilon\star(m_t^\varepsilon-m_*)(x)| \leqslant C_1e^{-ct}\,.$$

So for the first term in the definition (4.49) of ϕ, we have, for all $x\neq y$,

$$\begin{aligned}&\bigl|K^\varepsilon(x-y)\cdot\bigl(\nabla\ln m_t^\varepsilon(x)-\nabla\ln m_t^\varepsilon(y)\bigr)\bigr|\\&\quad=\biggl|K^\varepsilon(x-y)\cdot\biggl(\nabla\ln\frac{m_t^\varepsilon(x)}{m_*(x)}-\nabla\ln\frac{m_t^\varepsilon(y)}{m_*(y)}\biggr)\biggr|\\&\quad\lesssim\biggl\|\nabla^2\ln\frac{m_t^\varepsilon}{m_*}\biggr\|_{L^\infty}\leqslant\frac{C_1e^{-ct}}{\sqrt{t\wedge 1}}\,.\end{aligned}$$

For the second term, we have

$$K^\varepsilon \star m_t^\varepsilon(x) \cdot \nabla \ln m_t^\varepsilon(x) = K^\varepsilon \star (m^\varepsilon - m_*)(x) \cdot \nabla \ln m_* + K^\varepsilon \star m_t^\varepsilon(x) \cdot \nabla \ln \frac{m_t^\varepsilon(x)}{m_*(x)}\,,$$

therefore,

$$\begin{aligned}\|K^\varepsilon \star m_t^\varepsilon \cdot \nabla \ln m_t^\varepsilon\|_{L^\infty} &\leqslant \kappa_U \sup_x |x \cdot K^\varepsilon \star (m_t^\varepsilon - m_*)(x)| + \|K^\varepsilon \star m_t^\varepsilon\|_{L^\infty} \left\|\nabla \ln \frac{m_t^\varepsilon}{m_*}\right\|_{L^\infty} \\ &\leqslant C_2 e^{-ct}\,.\end{aligned}$$

Combining the two results above, we derive the decaying L^∞ bound for ϕ_t:

$$\|\phi_t\|_{L^\infty} \leqslant \frac{C_3 e^{-ct}}{\sqrt{t \wedge 1}}\,.$$

Thus, taking the alternative

$$\delta_t = \frac{3(1600^2 + 36e^4)C_3 e^{-ct}}{\sqrt{t \wedge 1}}\,,$$

we get

$$\frac{\mathrm{d}\mathcal{H}_t^\varepsilon}{\mathrm{d}t} \leqslant -\mathcal{I}_t^\varepsilon + \frac{C_4 e^{-ct}}{\sqrt{t \wedge 1}} \mathcal{H}_t^\varepsilon + \frac{C_4 e^{-ct}}{\sqrt{t \wedge 1}}\,.$$

Finally, we note that, as the Lipschitz constant of $\ln m_t^\varepsilon / m_*$ tends to zero exponentially, the perturbed measure m_t^ε satisfies a k_t-LSI with

$$k_t = 2\kappa_U \exp(-C_5 e^{-c' t})\,.$$

Thus, for all $t \geqslant 0$, we have

$$\mathcal{I}_t^\varepsilon \geqslant 2\kappa_U \exp(-C_5 e^{-c' t}) \mathcal{H}_t^\varepsilon\,.$$

We conclude by applying Grönwall's lemma and taking the limit $\varepsilon \to 0$. □

Chapter 5

Sharp local propagation of chaos for mean field particles with $W^{-1,\infty}$ kernels

Abstract. We present two methods to obtain $O(1/N^2)$ local propagation of chaos bounds for N diffusive particles in $W^{-1,\infty}$ mean field interaction. This extends the recent finding of Lacker [*Probab. Math. Phys.*, 4(2):377–432, 2023] to the case of singular interactions. The first method is based on a hierarchy of relative entropies and Fisher informations, and applies to the 2D viscous vortex model in the high temperature regime. Time-uniform local chaos bounds are also shown in this case. In the second method, we work on a hierarchy of L^2 distances and Dirichlet energies, and derive the desired sharp estimates for the same model in short time without restrictions on the temperature.

5.1 Introduction and main results

In this work, we are interested in the following system of $N \geqslant 2$ interacting particles on the d-dimensional torus $\mathbb{T}^d = (\mathbb{R}/\mathbb{Z})^d$:

$$\mathrm{d}X_t^i = \frac{1}{N-1} \sum_{j\in[N]:j\neq i} K\big(X_t^i - X_t^j\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i, \qquad \text{for } i \in [N], \tag{5.1}$$

where K is a singular interaction kernel, $W_\cdot^i$ are independent Brownian motions. and $[N] \coloneqq [\![1, N]\!] = \{1, \dots, N\}$. To be precise, we will consider kernels admitting the decomposition $K = K_1 + K_2$ such that K_1 is divergence-free and belongs to $W^{-1,\infty}(\mathbb{T}^d; \mathbb{R}^d)$, in the sense that $K_{1,\alpha} = \sum_{\beta=1}^d \partial_\beta V_{\beta\alpha}$ for some matrix field $V \in L^\infty(\mathbb{T}^d; \mathbb{R}^{d\times d})$, and $K_2 \in L^\infty(\mathbb{T}^d; \mathbb{R}^d)$. We then write the particle system's formal mean field limit when $N \to \infty$:

$$\mathrm{d}X_t = (K \star m_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \qquad m_t = \mathrm{Law}(X_t), \tag{5.2}$$

and wish to show that the system (5.1) converges to (5.2) when $N \to \infty$ in an appropriate sense.

The main example of the system in singular interaction is the *2D viscous vortex model*, where $d = 2$ and K is a periodic version of the following kernel defined on $\mathbb{R}^2$:

$$K'(x) = \frac{1}{2\pi}\frac{x^\perp}{|x|^2} = \frac{1}{2\pi}\left(-\frac{x_2}{|x|^2}, \frac{x_1}{|x|^2}\right)^\top, \qquad x = (x_1, x_2)^\top.$$

Notice that we have $K' = \nabla \cdot V'$ for

$$V'(x) = \frac{1}{2\pi}\begin{pmatrix} -\arctan(x_2/x_1) & 0 \\ 0 & \arctan(x_1/x_2) \end{pmatrix}.$$

The model originates from the studies of 2D incompressible Navier–Stokes equations and we refer readers to the work of Jabin and Z. Wang [124] and the expository article [205] (and references therein) for details.

Throughout the paper, we suppose that the N particles in the dynamics (5.1) are exchangeable, that is, for all permutation σ of the index set $[N]$, we have $\mathrm{Law}(X_t^1, \dots, X_t^N) = \mathrm{Law}(X_t^{\sigma(1)}, \dots, X_t^{\sigma(N)})$, and denote $m_t^{N,k} = \mathrm{Law}(X_t^1, \dots, X_t^k)$. The aim of this paper is then to investigate quantitatively the behavior of the distance between $m_t^{N,k}$ and $m_t^{\otimes k}$ when $N \to \infty$ and k remains fixed, that is, a quantitative *propagation of chaos* (PoC) phenomenon. The distances with which we work are the relative entropy

$$H(m_1|m_2) = \int \log\frac{m_1(x)}{m_2(x)} m_1(\mathrm{d}x)$$

and the so-called χ^2 distance

$$D(m_1|m_2) = \int \left(\frac{m_1(x)}{m_2(x)} - 1\right)^2 m_2(\mathrm{d}x)$$

The second distance will also be called the L^2 distance colloquially, if that leads to no confusion. In both of the two equations above, we have identified the probability laws m_1, m_2 with their density functions (with respect to the appropriate Lebesgue measure). The results of this paper are thus upper bounds on

$$H_t^k = H\big(m_t^{N,k}\big|m_t^{\otimes k}\big),\ D_t^k = D\big(m_t^{N,k}\big|m_t^{\otimes k}\big)$$

that are diminishing when $N \to \infty$. In the case of diffusion processes, the two crucial quantities

$$I(m_1|m_2) = \int \left|\nabla \log\frac{m_1(x)}{m_2(x)}\right|^2 m_1(\mathrm{d}x),$$
$$E(m_1|m_2) = \int \left|\nabla \frac{m_1(x)}{m_2(x)}\right|^2 m_2(\mathrm{d}x),$$

called respectively (relative) Fisher information and Dirichlet energy, also appear when we study the time-evolution of the relative entropy and the L^2 distance. In fact, the inclusion of these quantities in the analysis is the main novelty of this work.

Recently, the propagation of chaos phenomenon of singular mean field dynamics has raised high interests, and the main technique to overcome the singularity in the interaction is to study the evolution PDE describing the joint probability distribution of the N particles $m_t^N := m_t^{N,N} := \mathrm{Law}\big(X_t^1, \ldots, X_t^N\big)$, i.e. the *Liouville* or the *Fokker–Planck equation* of the particle system (5.1):

$$\partial_t m_t^N = \sum_{i\in[N]} \Delta_i m_t^N - \frac{1}{N-1} \sum_{i,j\in[N]:i\neq j} \nabla_i \cdot \big(m_t^N K(x^i - x^j)\big). \tag{5.3}$$

Notice that the N-tensorization $m_t^{\otimes N}$ of the mean field system (5.2) solves

$$\partial_t m_t^{\otimes N} = \sum_{i\in[N]} \Delta_i m_t^{\otimes N} - \sum_{i\in[N]} \nabla_i \cdot \big(m_t^{\otimes N} (K \star m_t)(x^i)\big). \tag{5.4}$$

Then it remains to find the appropriate functionals measuring the distance between m_t^N and $m_t^{\otimes N}$, and study the functionals' evolution. For $W^{-1,\infty}$ kernels with $W^{-1,\infty}$ divergences, Jabin and Z. Wang [124] have revealed that the relative entropy is the right functional and derived global-in-time PoC in this case.[1] For deterministic dynamics with repulsive or conservative Coulomb and Riesz interactions, Serfaty constructed the modulated energy in [209] and derived their global-in-time PoC. Then, Bresch, Jabin and Z. Wang [32, 31] extended the method of Serfaty to diffusive (and possibly attractive) Coulomb and Riesz systems and showed the global-in-time PoC by marrying relative entropy with modulated energy, the new functional being called modulated free energy. We mention here also another work [69] on the attractive case with logarithmic potentials. More recently, refinements of the methods above allow for uniform-in-time PoC estimates [98, 59] and extensions to the whole space have been done in [90, 201] and Chapter 4.

The main result of [124] applied to our dynamics (5.1), (5.2) already indicates

$$H\big(m_t^N \big| m_t^{\otimes N}\big) \leqslant Ce^{Ct}$$

for some $C \geqslant 0$, if the initial distance is zero: $m_0^N = m_0^{\otimes N}$. Then by the super-additivity of relative entropy, we get

$$H\big(m_t^{N,k} \big| m_t^{\otimes k}\big) \leqslant \frac{Ce^{Ct}}{\lfloor N/k \rfloor},$$

and this is already a quantitative PoC estimate. However, the findings of Lacker in [140] reveal that the $O(k/N)$-order bound obtained above is sub-optimal for regular interactions (where K is e.g. bounded), and the sharp order in this case is $O(k^2/N^2)$. The method of Lacker is to consider the BBGKY hierarchy of the marginal distrbutions $(m_t^{N,k})_{k\in[N]}$, where the evolution of $m_t^{N,k}$ depends on itself and the higher-level marginal $m_t^{N,k+1}$, namely

$$\begin{aligned}\partial_t m_t^{N,k} = \sum_{i\in[k]} \Delta_i m_t^{N,k} &- \frac{1}{N-1} \sum_{i,j\in[k]:i\neq j} \nabla_i \cdot \big(m_t^{N,k} K(x^i - x^j)\big) \\ &- \frac{N-k}{N-1} \sum_{i\in[k]} \nabla_i \cdot \bigg(\int_{\mathbb{T}^d} K(x^i - x_*) m_t^{N,k+1}(\boldsymbol{x}^{[k]}, x^*) \,\mathrm{d}x^*\bigg),\end{aligned} \tag{5.5}$$

and then to calculate the evolution of $H_t^k = H\big(m_t^{N,k} \big| m_t^{\otimes k}\big)$, which yields a hierarchy of ODE where $\mathrm{d}H_t^k/\mathrm{d}t$ depends on H_t^k and H_t^{k+1}. Solving this ODE system allows

[1] This work will be referred as "Jabin–Wang" in the following of this paper without including the name initial of the second author.

for the sharp $O(k^2/N^2)$ bounds on H_t^k. This method of Lacker is *local* in the sense that the quantity of interest describes the behavior of a fixed number of particles even when $N \to \infty$, and stand in contrast with the *global* approaches mentioned in the paragraph above, where the N-particle joint law is instead considered. Then, together with Le Flem, Lacker [142] strengthened his result and proved uniform-in-time $O(k^2/N^2)$ rate in a high temperature regime, with the help of log-Sobolev inequalities. Very recently, Hess-Childs and Rowan [111] extended this hierarchical method to the L^2 distance and obtained sharp convergence rates for higher-order expansions in the case of bounded interactions (the convergence of $m_t^{N,k}$ to the tensorized law $m_t^{\otimes k}$ being merely zeroth-order). One limitation of the entropy and L^2 methods is that we require the diffusivity of the dynamics to be non-zero, thus excluding deterministic Vlasov dynamics considered in the recent work of Duerinckx [78]. Still, two improvements are made possible via the entropy and L^2 methods. First, the norm-distance between $m_t^{N,k}$ and $m_t^{\otimes k}$ (which scales as the square root of relative entropy) can be shown to be of order $O(k/N)$, while directly applying the correlation bounds in [78] gives only an $O(k^2/N)$-order control. Note that this is also the order obtained in [182] for dynamics with collision terms. Second, we do not need to assume high regularity for the kernel and work with weaker norms for higher-order corrections as in [78], thanks to the fact that the Laplace operator prevents loss of derivatives in the BBGKY hierarchy. Finally, we note that Bresch, Jabin and coauthors have also applied hierarchical methods to study second-order dynamics of singular interaction in recent works [30, 29], and have shown respectively short-time strong PoC and global-in-time weak PoC under different regularity assumptions. This is significant progress, as the previous best PoC results for second-order systems, to the knowledge of the author, apply only to mildly singular kernels satisfying $K(x) = O(|x|^{-\alpha})$ for $\alpha < 1$.

In this work, we extend the entropic hierarchy of Lacker and the L^2 hierarchy of Hess-Childs–Rowan (only in the zeroth-order) to the case of $W^{-1,\infty}$ interactions. In the new hierarchies of ODE, which describe the evolution of H_t^k and D_t^k respectively, Fisher information and Dirichlet energy of the next level appear, and we develop new methods to solve the ODE systems. In the first entropic case, we show that $H_t^k = O(k^2/N^2)$ globally in time, if the temperature of the system is high enough (or equivalently, upon a rescaling of time, the interaction is weak enough). Moreover, in the case of 2D vortex model, we show that and $H_t^k = O(k^2 e^{-rt}/N^2)$ for some $r > 0$, thanks to the exponential decay established in [98, 59]. We also provide a simple way to solve Lacker's ODE system, based on a comparison principle. In the second L^2 case, we remove the restriction on the temperature by working with L^2 distances D_t^k and show that $D_t^k = O(1/N^2)$ for $k = O(1)$ but only in a short time interval.

We state the main results and discuss them in the rest of this section, and give their proof in Section 5.2. The studies of the ODE hierarchies, which are the final steps of the proof and the main technical contributions of this work, are postponed to Section 5.3. We present some other technical results in Section 5.4.

Throughout the paper, we will work with solution m_t^N of the Liouville equation (5.3) for which we can find a sequence of kernels $K^\varepsilon \in \mathcal{C}^\infty(\mathbb{T}^d)$ and probability densities $m_t^{N,\varepsilon} \in \mathcal{C}^\infty(\mathbb{T}^d)$ such that they satisfies (5.3) when K, m_t^N are respectively replaced by K^ε, $m_t^{N,\varepsilon}$; that $K^\varepsilon \to K$ almost everywhere and $m_t^{N,\varepsilon} \to m_t^N$ weakly as probability measures; and finally that $m_t^{N,\varepsilon}$ is lower bounded from 0. We suppose also that the mean field flow m_t is the weak limit of $\mathcal{C}^\infty$ approximations m_t^ε that correspond to the McKean–Vlasov SDE (5.2) driven by the regularized kernel K^ε, and that each m_t^ε has also strictly positive density. In particular, the 2D viscous vortex model verifies this assumption. See e.g. Chapter 4 for details. (Although the setting there is on $\mathbb{R}^d$ instead of $\mathbb{T}^d$ but the argument is the same.) We impose this technical assumption in order to avoid subtle well-posedness issues in the singular PDE (5.3) and we mention that it is also possible to work with entropy solutions for the same purpose. See [124] for details.

The main assumption of this paper is the following.

Assumption. The interaction kernel admits the decomposition $K = K_1 + K_2$, where $K_1 = \nabla \cdot V$ for some $V \in L^\infty(\mathbb{T}^d; \mathbb{R}^d \times \mathbb{R}^d)$ and satisfies $\nabla \cdot K_1 = 0$, and $K_2 \in L^\infty$.

We then state our main results.

Theorem 5.1 (Entropic PoC)**.** *Let the main assumption hold. Suppose that the marginal relative entropies at the initial time satisfy*

$$H_0^k \leqslant C_0 \frac{k^2}{N^2}$$

for all $k \in [N]$, for some $C_0 \geqslant 0$. If $\|V\|_{L^\infty} < 1$, then for all $T > 0$, there exists M, depending on

$$C_0,\ \|V\|_{L^\infty},\ \|K_2\|_{L^\infty},\ \sup_{t\in[0,T]} \|\nabla \log m_t\|_{L^\infty}^2 + \|\nabla^2 \log m_t\|_{L^\infty},$$

such that for all $t \in [0, T]$,

$$H_t^k \leqslant M e^{Mt} \frac{k^2}{N^2}.$$

If additionally $K_2 = 0$ and

$$\|\nabla \log m_t\|_{L^\infty}^2 + \|\nabla^2 \log m_t\|_{L^\infty} \leqslant M_m e^{-\eta t}$$

for all $t \geqslant 0$, for some $M_m \geqslant 0$ and $\eta > 0$, then for all r such that $0 < r < r_* \coloneqq \min\bigl(\eta, (1 - \|V\|_{L^\infty}) 8\pi^2\bigr)$, there exists M', depending on

$$C_0,\ \|V\|_{L^\infty},\ M_m,\ \eta,\ r,\ d,$$

such that for all $t \geqslant 0$, we have

$$H_t^k \leqslant M' e^{-rt} \frac{k^2}{N^2}.$$

Theorem 5.2 (L^2 PoC)**.** *Let the main assumption hold. Suppose that the marginal L^2 distances at the initial time satisfy*

$$D_0^k \leqslant C_0 \frac{k^2}{N^2}$$

for all $k \in [N]$, for some $C_0 \geqslant 0$. Let $T > 0$ be arbitrary. If the matrix field V satisfies

$$M_V \coloneqq \sup_{t\in[0,T]} \sup_{x\in\mathbb{T}^d} \int_{\mathbb{T}^d} |V(x-y)|^2 m_t(\mathrm{d}y) < 1,$$

then there exists $T_* > 0$, depending on

$$\|V\|_{L^\infty},\ M_V,\ \|K_2\|_{L^\infty},\ \sup_{t\in[0,T]} \|\nabla \log m_t\|^2_{L^\infty} + \|\nabla^2 \log m_t\|_{L^\infty},$$

such that for all $t \in [0, T_* \wedge T)$, we have

$$D^k_t \leqslant \frac{Me^{Mk}}{(T_* - t)^3 N^2}.$$

for some M depending additionally on C_0.

We discuss some consequences of the two theorems above.

$\nabla \cdot K_1 = 0$ is not restrictive. First, as noted in [124], the condition that the singular part K_1 is divergence-free is not restrictive. Indeed, if the interaction kernel K admits the decomposition $K = K'_1 + K'_2$, where both K'_1 and $\nabla \cdot K'_1$ belong to $W^{-1,\infty}$ (which is the regularity assumption of [124]), and $K'_2 \in L^\infty$, we can find, by definition, a bounded vector field S such that $\nabla \cdot K'_1 = \nabla \cdot S$. By shifting the components of S by constants, we can also suppose without loss of generality that this vector field verifies $\int_{\mathbb{T}^d} S = 0$. Thus, we have the alternative decomposition

$$K = (K'_1 - S) + (K'_2 + S),$$

where the first part $K'_1 - S$ is divergence-free and the second part $K'_2 + S$ is bounded. Since $S \in L^\infty$ and $\int_{\mathbb{T}^d} S = 0$, we can find a bounded matrix field V_S such that $\nabla \cdot V_S = S$ and $\|V_S\|_{L^\infty} \leqslant C_d \|S\|_{L^\infty}$ for some C_d depending only on the dimension d.[2] So the new decomposition satisfies the main assumption and it only remains to verify the respective "smallness" conditions of the two theorems for the kernel $K'_1 - S$.

2D vortex at high temperature. Second, Theorem 5.1 applies to the 2D viscous vortex model if the vortex interaction is weakly enough. Indeed, in the vortex case, we have $K = \nabla \cdot V$ for some $V \in L^\infty$ and $\nabla \cdot K = 0$ so the main assumption is satisfied with $K_2 = 0$. The required regularity bounds for the mean field flow m_t have been established in [98, 59]. More precisely, it is shown in [59, Section 3.2] that if the initial value m_0 of the mean field equation belongs to $W^{2,\infty}(\mathbb{T}^d)$ and verifies the lower bound $\inf m_0 > 0$, then we have the required decaying bound on the regularity:

$$\|\nabla \log m_t\|^2_{L^\infty} + \|\nabla^2 \log m_t\|_{L^\infty} \leqslant M_m e^{-\eta t}.\text{[3]}$$

[2] For example one can take $V^{1i}_S(x^1, x^2, \dots, x^d) = \int_0^{x^1} S^i(y, x^2, \dots, x^d)\,\mathrm{d}y$ for $i \in [d]$ and $V^{ji}_S = 0$ for $j \neq 1$.

[3] The rate of convergence stated in [59] is not explicit. However, it seems to the author that we can take $\eta = 4\pi^2$ by the following argument. First by computing the evolution of the entropy $H(m_t)$ and integrating by parts à la Jabin–Wang, we find that $\mathrm{d}H(m_t)/\mathrm{d}t = -I(m_t) \leqslant -8\pi^2 H(m_t)$ thanks to the log-Sobolev inequality (see also the proof of Theorem 4.24), and therefore $H(m_t) \lesssim e^{-8\pi^2 t}$. This implies that $\|m_t - 1\|_{L^1} \lesssim e^{-4\pi^2 t}$ by Pinsker. Then we use the hypercontractivity [59, Corollary 2.4] and the regularization [59, Proposition 2.6] to find that $\|\nabla m_t\|_{L^\infty}, \|\nabla^2 m_t\|_{L^\infty} \lesssim e^{-4\pi^2 t}$ so the desired bound follows with $\eta = 4\pi^2$. This rate is optimal as it is verified by the heat equation ($K = 0$) with initial data $m_0(x) = 1 + a\sin(2\pi x) + b\cos(2\pi x)$. With $\eta = 4\pi^2$, the minimum for the rate in the second assertion of Theorem 5.1 is equal to $\min(1, 2 - 2\|V\|_{L^\infty})4\pi^2$.

So Theorem 5.1 applies if $\|V\|_{L^\infty} < 1$. Upon a time-rescaling, this result can be extended to 2D viscous vortex at any temperature $\tau > 0$ (where the diffusion coefficient in (5.1) is $\sqrt{\tau}$ instead of $\sqrt{2}$), once $\|V\|_{L^\infty} < \tau/2$. In this high temperature regime, the second assertion of Theorem 5.1 provides a finer long-time convergence estimate on the relative entropies for the 2D viscous vortex model compared to the global results in [98, 59]. These results seems to be new, but it is unclear to the author if the high-temperature restriction can be lifted. (See also the discussion on L^2 results in below.)

L^d interaction at any temperature. On the contrary, if the interaction kernel K is of the slightly higher regularity class

$$K \in L^d,\ \nabla \cdot K \in L^d,$$

then Theorem 5.1 can be applied without any restriction on the strengh of K. To this end, we consider $K^\varepsilon = K \star \rho^\varepsilon$ where ρ^ε is a sequence of $\mathcal{C}^\infty$ mollifiers on $\mathbb{T}^d$. Since $\int_{\mathbb{T}^d} K - K^\varepsilon = 0$ and $\int_{\mathbb{T}^d} \nabla \cdot K - \nabla \cdot K^\varepsilon = 0$, the result of Bourgain and Brezis [28] indicates that we can find a matrix field V and a vector field S on $\mathbb{T}^d$ solving the equations $\nabla \cdot V = K - K^\varepsilon$ and $\nabla \cdot S = \nabla \cdot K - \nabla \cdot K^\varepsilon$ with the bounds

$$\begin{aligned}\|V\|_{L^\infty} &\leqslant C_d \|K - K^\varepsilon\|_{L^\infty},\\ \|S\|_{L^\infty} &\leqslant C_d \|\nabla \cdot K - \nabla \cdot K^\varepsilon\|_{L^\infty}\end{aligned}$$

for some $C_d > 0$ depending only on d. By shifting the components of S, we can suppose that $\int_{\mathbb{T}^d} S = 0$ and this does not alter the L^∞ bound on S above. We find again a matrix field V_S such that $\nabla \cdot V_S = S$ and $\|V_S\|_{L^\infty} \leqslant C_d \|S\|_{L^\infty}$. Then we decompose the kernel K in the following way:

$$K = (K - K^\varepsilon) + K^\varepsilon = \nabla \cdot V + K^\varepsilon = \nabla \cdot (V - V_S) + (K^\varepsilon + S).$$

By construction, the singular part is divergence-free:

$$\nabla^2 : (V - V_S) = \nabla \cdot (K - K^\varepsilon) - \nabla \cdot S = 0,$$

and the remaining part $K^\varepsilon + S$ is bounded, so the main assumption is satisfied. The $W^{-1,\infty}$ norm of the singular part is controlled by

$$\|V - V_S\|_{L^\infty} \leqslant \|V\|_{L^\infty} + \|V_S\|_{L^\infty} \leqslant C_d \big(\|K - K^\varepsilon\|_{L^d} + \|\nabla \cdot K - \nabla \cdot K^\varepsilon\|_{L^d}\big).$$

Yet, the mollification is continuous in L^d:

$$\|K - K^\varepsilon\|_{L^d},\ \|\nabla \cdot K - \nabla \cdot K^\varepsilon\|_{L^d} \to 0, \qquad \text{when } \varepsilon \to 0.$$

So in order to apply Theorem 5.1, it suffices to take an ε small enough. In a previous work, Han [105, Theorem 1.2] derived global $O(1/N^2)$ PoC under the assumption that K is divergence-free and belongs to L^p for some $p > d$, and the N-particle initial measure satisfies the density bound $\lambda^{-1} \leqslant m_0^N \leqslant \lambda$ uniformly in N. In comparison to this work, our method achieves two major improvements: first, the critical Krylov–Röckner exponent $p = d$ is treated [136]; and second, the rather demanding condition on m_0^N (which excludes non-trivial chaotic data $m_0^N = m_0^{\otimes N}$ for $m_0 \neq 1$) is lifted. These improvements are made possible by our consideration of the new hierarchy involving Fisher information (see Proposition 5.5) and a Jabin–Wang type large deviation estimate (see Corollary 5.10).

2D vortex at any temperature through L^2. By a similar regularity trick, the L^2 result of Theorem 5.2 can be applied to the 2D viscous vortex model at any temperature (or equivalently, without restriction on the interaction strength). Indeed, as in the case, $K = \nabla \cdot V$ for $V \in L^\infty$ and $\nabla \cdot K = 0$, we can decompose

$$K = (K - K^\varepsilon) + K^\varepsilon = \nabla \cdot (V - V^\varepsilon) + K^\varepsilon,$$

where $K^\varepsilon = K \star \rho^\varepsilon$ and $V^\varepsilon = V \star \rho^\varepsilon$. Then the L^2 constant in Theorem 5.2 satisfies

$$M_{V-V^\varepsilon} \coloneqq \sup_{t\in[0,T]} \sup_{x\in\mathbb{T}^d} \int_{\mathbb{T}^d} |(V - V^\varepsilon)(x-y)|^2 m_t(\mathrm{d}y) \leqslant \|V - V^\varepsilon\|^2_{L^2} \sup_{t\in[0,T]} \|m_t\|_{L^\infty},$$

and can be arbitrarily small as $\varepsilon \to 0$. Thus Theorem 5.2 gives an $O(1/N^2)$ PoC estimate in short time. Since our treatment of the L^2 hierarchy in Proposition 5.6 is rather crude, it seems possible to the author that the explosion in finite time is sub-optimal. Here, the major technical difficulty is that we cannot force the hierarchy to stop at a certain level $k \sim N^\alpha$, $\alpha < 1$ as done in Hess-Child–Rowan [111]. And this is due to the fact that we do not have a priori bounds on L^2 distances and Dirichlet energies that are strong enough.

Dynamics on the whole space. As a concluding remark, we could also expect that similar results on $O(1/N^2)$ PoC hold for dynamics on the whole space, since the Jabin–Wang results have been migrated to that case ([90, 201] and Chapter 4), and the original theorem of Lacker [140] is already on $\mathbb{R}^d$.

5.2 Proof of Theorems 5.1 and 5.2

5.2.1 Setup and proof outline

In the proof we will work with regularized solutions introduced in Section 5.1 and prove the bounds in both theorems for these approximations. Then the result holds for the original solutions by lower semi-continuity. See Chapter 4 for details.

In the following, we will perform the entropic and L^2 computations at the same time in order to exploit the similarity between them. We set $p = 1$ for the entropic computations and $p = 2$ for the L^2 computations. Then, we can write the relative entropy and the L^2 distance between $m_t^{N,k}$ and $m_t^{\otimes k}$ formally as

$$\mathcal{D}_p^k \coloneqq \mathcal{D}_p\big(m_t^{N,k} \big| m_t^{\otimes k}\big) \coloneqq \frac{1}{p-1}\bigg(\int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^p \,\mathrm{d}m_t^{\otimes k} - 1\bigg), \quad \text{where } h_t^{N,k} \coloneqq \frac{m_t^{N,k}}{m_t^{\otimes k}}.$$

The expression makes sense classically in the L^2 case where $p = 2$. In the entropic case, this notation is motivated by the fact that

$$\lim_{p\searrow 1} \frac{1}{p-1}\bigg(\int h^p \,\mathrm{d}m - 1\bigg) = \int h \log h \,\mathrm{d}m$$

for all postive h that is upper and lower bounded (away from zero) and all probability measure m such that $\int h \,\mathrm{d}m = 1$.

Then, we use the BBGKY hierarchy (5.5) and the tensorized mean field equation (5.4) to calculate the time derivative of $\mathcal{D}_p^k$. We find

$$\begin{aligned}\frac{1}{p}\frac{\mathrm{d}\mathcal{D}_p^k}{\mathrm{d}t} = &-\int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^{p-2}\big|\nabla h_t^{N,k}\big|^2 \,\mathrm{d}m_t^{\otimes k} \\ &+ \frac{1}{N-1}\sum_{i,j\in[k]:i\neq j}\int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^{p-1}\nabla_i h_t^{N,k} \\ &\qquad\qquad \cdot \big(K(x^i-x^j) - K\star m_t(x^i)\big)m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &+ \frac{N-k}{N-1}\sum_{i\in[k]}\int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^{p-1}\nabla_i h_t^{N,k} \\ &\qquad\qquad \cdot \Big\langle K(x^i-\cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t\Big\rangle m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}),\end{aligned}$$

where the conditional measure $m_t^{N,(k+1)|k}(\cdot|\cdot)$ is defined as

$$m_t^{N,(k+1)|k}(x^*|\boldsymbol{x}^{[k]}) \coloneqq \frac{m_t^{N,k+1}(\boldsymbol{x}^{[k]},x^*)}{m_t^{N,k}(\boldsymbol{x}^{[k]})}.$$

Define also

$$\mathcal{E}_p^k \coloneqq \int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^{p-2}\big|\nabla h_t^{N,k}\big|^2 \,\mathrm{d}m_t^{\otimes k}.$$

This expression makes sense for both $p=1$ and 2, and is the relative Fisher information $I_t^k = I\big(m_t^{N,k}\big|m_t^{\otimes k}\big)$ for $p=1$, and the Dirichlet energy $E_t^k = E\big(m_t^{N,k}\big|m_t^{\otimes k}\big)$ for $p=2$. Denote by A and B the last two terms in the equality above for $p^{-1}\,\mathrm{d}D_p^k/\mathrm{d}t$. We find that $A = A_1 + A_2$ and $B = B_1 + B_2$ where

$$A_a \coloneqq \frac{1}{N-1}\sum_{i,j\in[k]:i\neq j}\int_{\mathbb{T}^{dk}} \big(h_t^{N,k}\big)^{p-1}\nabla_i h_t^{N,k}\cdot\big(K_a(x^i-x^j)-K_a\star m_t(x^i)\big)m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})$$

and

$$\begin{aligned}B_a \coloneqq \frac{N-k}{N-1}&\sum_{i\in[k]}\int_{\mathbb{T}^{dk}} \big(h_t^{N,k}\big)^{p-1}\nabla_i h_t^{N,k} \\ &\cdot \Big\langle K_a(x^i-\cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t\Big\rangle m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}),\end{aligned}$$

for $a=1$, 2, since the expressions are linear in K and the kernel admits the decomposition $K = K_1 + K_2$. Thus, the evolution of $\mathcal{D}_p^k$ writes

$$\frac{1}{p}\frac{\mathrm{d}\mathcal{D}_p^k}{\mathrm{d}t} = -\mathcal{E}_p^k + A_1 + A_2 + B_1 + B_2.$$

We call A_1, A_2 the *inner interaction* terms, and B_1, B_2 the *outer interaction* terms, as the first two terms correspond to the interaction between the first k particles themselves, and the last two terms to the interaction between the first k and the remaining $N-k$ particles.

We aim to find appropriate upper bounds for the last four interaction terms A_1, A_2, B_1, B_2 in the rest of the proof. To be precise, we will show in the entropic case $p=1$ the following system of differential inequalities:

$$\frac{\mathrm{d}H_t^k}{\mathrm{d}t} \leqslant -c_1 I_t^k + c_2 I_t^{k+1}\mathbb{1}_{k<N} + M_1 H_t^k + M_2 k\big(H_t^{k+1} - H_t^k\big)\mathbb{1}_{k<N} + M_3\frac{k^\beta}{N^2},$$

where β is an integer $\geqslant 2$ and c_1, c_2, M_i, $i \in [3]$ are nonnegative constants such that $c_1 > c_2$. And in the L^2 case $p = 2$, we show that

$$\frac{\mathrm{d}D_t^k}{\mathrm{d}t} \leqslant -c_1 E_t^k + c_2 E_t^{k+1} \mathbb{1}_{k<N} + M_2 k D_t^{k+1} \mathbb{1}_{k<N} + M_3 \frac{k^2}{N^2},$$

where again $c_1 > c_2 \geqslant 0$ and M_2, $M_3 \geqslant 0$. We will then apply the results from the following section (Propositions 5.5 and 5.6) to solve the hierarchies and this will conclude the proof.

5.2.2 Two lemmas on inner interaction terms

We establish two lemmas that will be useful for controlling the inner interactions terms A_1, A_2.

Lemma 5.3. *Let $p \in \{1, 2\}$ and k be an integer $\geqslant 2$. Let $m \in \mathcal{P}(\mathbb{T}^d)$ and $h : \mathbb{T}^{kd} \to \mathbb{R}_{\geqslant 0}$ be exchangeable. Suppose additionally that $\int_{\mathbb{T}^{kd}} h \, \mathrm{d}m^{\otimes k} = 1$. Let $U : \mathbb{T}^{2d} \to \mathbb{R}^d$ be bounded. For $i \in [k]$, denote*

$$a := \sum_{j \in [k] : j \neq i} \int_{\mathbb{T}^{kd}} h^{p-1} \nabla_i h \cdot \big(U(x^i, x^j) - \langle U(x^i, \cdot), m \rangle \big) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}),$$

where $\langle U(x^i, \cdot), m \rangle = \int_{\mathbb{T}^d} U(x^i, y) m(\mathrm{d}y)$. Then in the case $p = 1$, we have for all $\varepsilon > 0$,[4]

$$a \leqslant \varepsilon \int_{\mathbb{T}^{kd}} \frac{|\nabla_i h|^2}{h} \, \mathrm{d}m^{\otimes k} + \frac{\|U\|_{L^\infty}^2}{\varepsilon} \times \begin{cases} (k-1)^2 \\ (k-1) + (k-1)(k-2)\sqrt{2H(m^3 | m^{\otimes 3})} \end{cases}$$

where m^3 is the 3-marginal of the probability measure $hm^{\otimes k}$:

$$m^3(\mathrm{d}x^{[3]}) = \int_{\mathbb{T}^{(k-3)d}} h m^{\otimes k} \, \mathrm{d}\boldsymbol{x}^{[k] \setminus [3]}.$$

And in the case $p = 2$, we have for all $\varepsilon > 0$,

$$a \leqslant \varepsilon \int_{\mathbb{T}^{kd}} |\nabla_i h|^2 \, \mathrm{d}m^{\otimes k} + \frac{2(k-1)^2 \|U\|_{L^\infty}^2}{\varepsilon} D + \frac{2(k-1)\|U\|_{L^\infty}^2}{\varepsilon},$$

where $D = \int_{\mathbb{T}^{kd}} (h-1)^2 \, \mathrm{d}m^{\otimes k}$.

Proof of Lemma 5.3. This estimate with $p = 1$ has already been established in [140], and with $p = 2$ it is done implicitly in [111]. Nevertheless, we give a full proof here for self-containedness. In the simpler case $p = 2$, using the Cauchy–Schwarz inequality

$$h \nabla_i h \cdot \xi = \big((h-1) + 1 \big) \nabla_i h \cdot \xi \leqslant \varepsilon |\nabla_i h|^2 + \frac{1}{2\varepsilon} \big((h-1)^2 + 1 \big) |\xi|^2,$$

we get

$$\sum_{j \in [k] : j \neq i} h \nabla_i h \cdot \big(U(x^i, x^j) - \langle U(x^i, \cdot), m \rangle \big)$$
$$\leqslant \varepsilon |\nabla_i h|^2 + \frac{1}{2\varepsilon} \big((h-1)^2 + 1 \big) \bigg| \sum_{j \in [k] : j \neq i} \big(U(x^i, x^j) - \langle U(x^i, \cdot), m \rangle \big) \bigg|^2$$

[4]Here, and in the following, if a bracket without conditions appears in a math expression, it means that both alternatives are valid.

Thus, integrating against $m^{\otimes k}$, we get

$$\begin{aligned}
&\sum_{j\in[k]:j\neq i}\int_{\mathbb{T}^{kd}} h^{p-1}\nabla_i h\cdot\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\big)m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})\\
&\quad\leqslant \varepsilon\int_{\mathbb{T}^{kd}}|\nabla_i h|^2\,\mathrm{d}m^{\otimes k}\\
&\qquad+\frac{1}{2\varepsilon}\int_{\mathbb{T}^{kd}}\big((h-1)^2+1\big)\Big|\sum_{j\in[k]:j\neq i}\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\big)\Big|^2 m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})\\
&\quad\leqslant \varepsilon\int_{\mathbb{T}^{kd}}|\nabla_i h|^2\,\mathrm{d}\boldsymbol{x}^{[k]}+\frac{(k-1)^2\|U\|_{L^\infty}^2}{2\varepsilon}D\\
&\qquad+\frac{1}{2\varepsilon}\int_{\mathbb{T}^{kd}}\Big|\sum_{j\in[k]:j\neq i}\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\Big|^2 m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}).
\end{aligned}$$

The integral in the last term is equal to

$$\sum_{j_1,j_2\in[k]\setminus\{i\}}\int_{\mathbb{T}^{kd}}\big(U(x^i,x^{j_1})-\langle U(x^i,\cdot),m\rangle\big)\cdot\big(U(x^i,x^{j_2})-\langle U(x^i,\cdot),m\rangle\big)m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}),$$

and we notice that by independence, the integral above does not vanish only if $j_1=j_2$. Thus we get the upper bound

$$\int_{\mathbb{T}^{kd}}\Big|\sum_{j\in[k]:j\neq i}\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle(x^i)\Big|^2 m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})\leqslant 4(k-1)\|U\|_{L^\infty}^2,$$

and this finishes the proof for the $p=2$ case.

Now treat the entropic case where $p=1$. Using Cauchy–Schwarz, we get

$$\begin{aligned}
&\sum_{j\in[k]:j\neq i}\nabla_i h\cdot\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\big)\\
&\qquad\qquad\leqslant \varepsilon h^{-1}|\nabla_i h|^2+\frac{1}{4\varepsilon}\Big|\sum_{j\in[k]:j\neq i}\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\big)\Big|^2.
\end{aligned}$$

Then integrating against $m^{\otimes k}$, we find

$$\begin{aligned}
&\sum_{i,j\in[k]:j\neq i}\int_{\mathbb{T}^{kd}}\nabla_i h\cdot\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\big)m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})\\
&\quad\leqslant \varepsilon\int_{\mathbb{T}^{kd}}\frac{|\nabla_i h|^2}{h}\,\mathrm{d}m^{\otimes k}\\
&\qquad+\frac{1}{4\varepsilon}\int_{\mathbb{T}^{kd}}\Big|\sum_{j\in[k]:j\neq i}\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\big)\Big|^2 hm^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}).
\end{aligned}$$

So it remains to upper bound the last integral. Employing the crude bound

$$\Big|\sum_{j\in[k]:j\neq i}\big(U(x^i,x^j)-\langle U(x^i,\cdot),m\rangle\big)\Big|^2\leqslant 4(k-1)^2\|U\|_{L^\infty}^2$$

and the fact that $hm^{\otimes k}$ is a probability measure, we get

$$\int_{\mathbb{T}^{kd}} \Big| \sum_{j\in[k]:j\neq i} \big(U(x^i,x^j) - \langle U(x^i,\cdot),m\rangle\big)\Big|^2 hm^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \leqslant 4(k-1)^2\|U\|_{L^\infty}^2.$$

This yields the first claim for the case $p=1$. For the finer bound, we again expand the square in the integrand:

$$\begin{aligned}
&\int_{\mathbb{T}^{kd}} \Big| \sum_{j\in[k]:j\neq i} \big(U(x^i,x^j) - \langle U(x^i,\cdot),m\rangle\big)\Big|^2 hm^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\
&\quad = \sum_{j\in[k]\setminus\{i\}} \int_{\mathbb{T}^{kd}} |U(x^i,x^j) - \langle U(x^i,\cdot),m\rangle|^2 hm^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\
&\qquad + \sum_{j_1,j_2\in[k]\setminus\{i\}:j_1\neq j_2} \int_{\mathbb{T}^{kd}} \big(U(x^i,x^{j_1}) - \langle U(x^i,\cdot),m\rangle\big) \\
&\qquad\qquad \cdot \big(U(x^i,x^{j_2}) - \langle U(x^i,\cdot),m\rangle\big) hm^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}).
\end{aligned}$$

The first term can be bounded crudely by $4(k-1)\|U\|_{L^\infty}^2$ as before. For the second term, we notice that the integration against the measure $hm^{\otimes k}$ can be replaced by the integration against the 3-marginal

$$m^3(\mathrm{d}x^i\,\mathrm{d}x^{j_1}\,\mathrm{d}x^{j_2}) = \int_{\mathbb{T}^{(k-3)d}} hm^{\otimes k}\,\mathrm{d}\boldsymbol{x}^{[k]\setminus\{i,j_1,j_2\}}.$$

Notice that, by independence, we have

$$\int_{\mathbb{T}^{3d}} \big(U(x^i,x^{j_1}) - \langle U(x^i,\cdot),m\rangle\big)\cdot\big(U(x^i,x^{j_2}) - \langle U(x^i,\cdot),m\rangle\big) m^{\otimes 3}(\mathrm{d}x^i\,\mathrm{d}x^{j_1}\,\mathrm{d}x^{j_2}) = 0.$$

Using the Pinsker inequality between m^3 and $m^{\otimes 3}$, we find for $j_1\neq j_2$,

$$\begin{aligned}
&\int_{\mathbb{T}^{3d}} \big(U(x^i,x^{j_1}) - \langle U(x^i,\cdot),m\rangle\big)\cdot\big(U(x^i,x^{j_2}) - \langle U(x^i,\cdot),m\rangle\big) m^{3}(\mathrm{d}x^i\,\mathrm{d}x^{j_1}\,\mathrm{d}x^{j_2}) \\
&\qquad \leqslant 4\|U\|_{L^\infty}^2\sqrt{2H(m^3|m^{\otimes 3})},
\end{aligned}$$

and this concludes the proof for the case $p=1$. □

Lemma 5.4. *Under the same setting as in Lemma 5.3, let $\phi:\mathbb{T}^{2d}\to\mathbb{R}$ be a bounded function verifying $\phi(x,x)=0$ for all $x\in\mathbb{T}^d$ and the (second-order) cumulant property:*

$$\int_{\mathbb{T}^d}\phi(x,y)m(\mathrm{d}y) = \int_{\mathbb{T}^d}\phi(y,x)m(\mathrm{d}x) = 0, \qquad \text{for all } x\in\mathbb{T}^d.$$

Then we have

$$\begin{aligned}
&\sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} h^p\phi(x^i,x^j)m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\
&\qquad \leqslant \|\phi\|_{L^\infty}\bigg[\sqrt{2C_{\mathrm{JW}}}N\Big(\mathcal{D}_p + \frac{3k^2}{N^2}\Big) + k^2\mathcal{D}_p\mathbb{1}_{p=2}\bigg],
\end{aligned}$$

where C_{JW} is a universal constant to be defined in Section 5.4.2 and $\mathcal{D}_p$ is defined by

$$\mathcal{D}_p := \begin{cases} \int_{\mathbb{T}^{kd}} h \log h \,\mathrm{d}m^{\otimes k} & \text{when } p = 1, \\ \int_{\mathbb{T}^{kd}} (h-1)^2 \,\mathrm{d}m^{\otimes k} & \text{when } p = 2. \end{cases}$$

Proof of Lemma 5.4. In the case $p = 1$, thanks to the convex duality of entropy, we have

$$\begin{aligned} &\sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} h\phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\quad = \sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} (h-1)\phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\quad \leqslant \eta^{-1} \int_{\mathbb{T}^{kd}} h \log h \,\mathrm{d}m^{\otimes k} + \eta^{-1} \log \int_{\mathbb{T}^{kd}} \exp\biggl(\eta \sum_{i,j\in[k]} \phi(x^i, x^j)\biggr) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}), \end{aligned}$$

for all $\eta > 0$. Then taking η such that $\sqrt{2C_{\mathrm{JW}}} \|\phi\|_{L^\infty} N\eta = 1$ and applying the modified Jabin–Wang estimates in Corollary 5.10, we get

$$\sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} h\phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \leqslant \sqrt{2C_{\mathrm{JW}}} \|\phi\|_{L^\infty} N \biggl(\mathcal{D}_1 + \frac{3k^2}{N^2}\biggr).$$

In the case $p = 2$, we use the elementary equality

$$h^2 = (h-1)^2 + 2(h-1) + 1$$

and get

$$\begin{aligned} &\sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} h^2\phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\quad = \sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} (h-1)^2\phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\qquad\quad + 2 \sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} (h-1)\phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\quad \leqslant k^2 \|\phi\|_{L^\infty} \int_{\mathbb{T}^{kd}} (h-1)^2 \,\mathrm{d}m^{\otimes k} \\ &\qquad\quad + 2\biggl(\int_{\mathbb{T}^{kd}} (h-1)^2 \,\mathrm{d}m^{\otimes k}\biggr)^{1/2} \biggl[\int_{\mathbb{T}^{kd}} \biggl(\sum_{i,j\in[k]} \phi(x^i, x^j)\biggr)^2 \mathrm{d}m^{\otimes k}\biggr]^{1/2} \end{aligned}$$

The last integral has already been estimated in the intermediate (and in fact the easiest) step of the Jabin–Wang large deviation lemma (see Proposition 5.9):

$$\int_{\mathbb{T}^{kd}} \biggl(\sum_{i,j\in[k]} \phi(x^i, x^j)\biggr)^2 \mathrm{d}m^{\otimes k} \leqslant 2k^2 C_{\mathrm{JW}} \|\phi\|_{L^\infty}^2.$$

Thus we have

$$\int_{\mathbb{T}^{kd}} h^2\phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \leqslant k^2 \|\phi\|_{L^\infty} \mathcal{D}_2 + 2k\|\phi\|_{L^\infty} \sqrt{2C_{\mathrm{JW}}\mathcal{D}_2},$$

so the desired result follows from the Cauchy–Schwarz inequality. □

5.2.3 Control of the inner interaction terms

In this step, we aim to find appropriate upper bounds for the inner interactions terms

$$A_a := \frac{1}{N-1}\sum_{i,j\in[k]:i\neq j}\int_{\mathbb{T}^{dk}}\big(h_t^{N,k}\big)^{p-1}\nabla_i h_t^{N,k}\cdot\big(K_a(x^i-x^j)-K_a\star m_t(x^i)\big)m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}),$$

where $p = 1$, 2 and $a = 1$, 2.

Control of the regular part A_2

First start with the regular part. In this case, we directly invoke Lemma 5.3 with $U(x,y) = K_2(x-y)$ and $\varepsilon = (N-1)\varepsilon_1$ for some $\varepsilon_1 > 0$. Summing over $i \in [k]$, we get

$$A_2 \leqslant \varepsilon_1 I_t^k + \frac{C\|K_2\|_{L^\infty}^2 k}{\varepsilon_1(N-1)^2}\times\begin{cases}(k-1)^2\\(k-1)+(k-1)(k-2)\sqrt{H_t^3}\end{cases}$$

for the case $p = 1$, and

$$A_2 \leqslant \varepsilon_1 E_t^k + \frac{C\|K_2\|_{L^\infty}^2 k(k-1)^2}{\varepsilon_1(N-1)^2}D_t^k + \frac{C\|K_2\|_{L^\infty}^2 k(k-1)}{\varepsilon_1(N-1)^2}$$

for the case $p = 2$. In both inequalities above, C denotes a universal constant that may change from line to line, and we adopt this convention in the rest of the proof.

Control of the singular part A_1

Recall that $K_1 = \nabla\cdot V$ and $\nabla\cdot K_1 = 0$. Then we perform the integrations by parts:

$$\begin{aligned}
&p(N-1)A_1\\
&\quad= p\sum_{i,j\in[k]:i\neq j}\int_{\mathbb{T}^{kd}}\big(h_t^{N,k}\big)^{p-1}\nabla_i h_t^{N,k}\cdot\big(K_1(x^i-x^j)-(K_1\star m_t)(x^i)\big)m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})\\
&\quad= \sum_{i,j\in[k]:i\neq j}\int_{\mathbb{T}^{kd}}\nabla_i\big(h_t^{N,k}\big)^{p}\cdot\big(K_1(x^i-x^j)-(K_1\star m_t)(x^i)\big)m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})\\
&\quad= -\sum_{i,j\in[k]:i\neq j}\int_{\mathbb{T}^{kd}}\big(h_t^{N,k}\big)^{p}\nabla\log m_t(x^i)\\
&\qquad\qquad\qquad\cdot\big(K_1(x^i-x^j)-(K_1\star m_t)(x^i)\big)m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})\\
&\quad= \sum_{i,j\in[k]:i\neq j}\int_{\mathbb{T}^{kd}}\nabla_i\Big(\big(h_t^{N,k}\big)^{p}\nabla\log m_t(x^i)m_t^{\otimes k}\Big)\\
&\qquad\qquad\qquad:\big(V(x^i-x^j)-(V\star m_t)(x^i)\big)\,\mathrm{d}\boldsymbol{x}^{[k]}.
\end{aligned}$$

Noticing that $\nabla\log m_t(x^i)m_t^{\otimes k} = \nabla_i\big(m_t^{\otimes k}\big)$, we get

$$\begin{aligned}
&\nabla_i\Big(\big(h_t^{N,k}\big)^{p}\nabla\log m_t(x^i)m_t^{\otimes k}\Big)\\
&\qquad= p\big(h_t^{N,k}\big)^{p-1}\nabla_i h_t^{N,k}\otimes\nabla\log m_t(x^i)m_t^{\otimes k} + \big(h_t^{N,k}\big)^{p}\frac{\nabla^2 m_t(x^i)}{m_t(x^i)}m_t^{\otimes k}.
\end{aligned}$$

Hence,

$$
\begin{aligned}
&p(N-1)A_1 \\
&\quad = p \sum_{i,j\in[k]:i\neq j} \int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^{p-1} \nabla_i h_t^{N,k} \otimes \nabla \log m_t(x^i) \\
&\qquad\qquad\qquad\qquad\qquad : \big(V(x^i - x^j) - (V \star m_t)(x^i)\big) m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\
&\qquad + \sum_{i,j\in[k]:i\neq j} \int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^{p} \frac{\nabla^2 m_t(x^i)}{m_t(x^i)} : \big(V(x^i - x^j) - (V \star m_t)(x^i)\big) m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\
&\quad =: p(N-1)(A_{11} + A_{12}).
\end{aligned}
$$

For the first part A_{11}, we invoke Lemma 5.3 with $U(x,y) = \nabla \log m_t(x) \cdot V(x-y)$ and $\varepsilon = (N-1)\varepsilon_2$ for some $\varepsilon_2 > 0$. Summing over $i \in [k]$, we get

$$
A_{11} \leqslant \varepsilon_2 I_t^k + \frac{C\|\nabla \log m_t\|_{L^\infty}^2 \|V\|_{L^\infty}^2 k}{\varepsilon_2 (N-1)^2} \times \begin{cases} (k-1)^2 \\ (k-1) + (k-1)(k-2)\sqrt{H_t^3} \end{cases}
$$

for the case $p = 1$, and

$$
A_{11} \leqslant \varepsilon_2 E_t^k + \frac{C\|\nabla \log m_t\|_{L^\infty}^2 \|V\|_{L^\infty}^2 k(k-1)^2}{\varepsilon_2 (N-1)^2} D_t^k + \frac{C\|\nabla \log m_t\|_{L^\infty}^2 \|V\|_{L^\infty}^2 k(k-1)}{\varepsilon_2 (N-1)^2}
$$

for the case $p = 2$.

For the second part A_{12}, we invoke Lemma 5.4 with

$$
\phi(x,y) = \begin{cases} \frac{\nabla^2 m_t(x)}{m_t(x)} : \big(V(x-y) - (V \star m_t)(x)\big) & \text{if } x \neq y, \\ 0 & \text{if } x = y. \end{cases}
$$

Note that the cumulant condition

$$
\int_{\mathbb{T}^d} \phi(x,y) m_t(\mathrm{d}y) = \int_{\mathbb{T}^d} \phi(y,x) m_t(\mathrm{d}y) = 0
$$

is verified due to the definition of convolution and the fact that $\nabla^2 : V = \nabla \cdot K_1 = 0$. Thus, we get

$$
A_{12} \leqslant \frac{\|\nabla^2 m_t / m_t\|_{L^\infty} \|V\|_{L^\infty}}{N-1} \left[CN\left(\mathcal{D}_p^k + \frac{k^2}{N^2} \right) + k^2 \mathcal{D}_p^k \mathbb{1}_{p=2} \right]
$$

where C is a universal constant.

Denote

$$
M_{V,m_t} \coloneqq \|\nabla \log m_t\|_{L^\infty}^2 \|V\|_{L^\infty}^2 + \|\nabla^2 m_t / m_t\|_{L^\infty} \|V\|_{L^\infty},
$$

and note that here, since $\nabla^2 m_t / m_t = (\nabla \log m_t)^{\otimes 2} + \nabla^2 \log m_t$, the constant M_{V,m_t} is finite by the assumptions of the theorems. Summing up A_{11} and A_{12}, we get

$$
A_1 \leqslant \varepsilon_2 I_t^k + CM_{V,m_t}\left(H_t^k + \frac{k^2}{N^2} \right) + \frac{CM_{V,m_t} k}{\varepsilon_2 N^2} \times \begin{cases} k^2 \\ k + k^2 \sqrt{H_t^3} \end{cases}
$$

for the case $p = 1$, and

$$
A_1 \leqslant \varepsilon_2 E_t^k + CM_{V,m_t}\left(1 + \frac{k^2}{N} + \frac{k^3}{\varepsilon_2 N^2} \right) D_t^k + CM_{V,m_t}(1 + \varepsilon_2^{-1}) \frac{k^2}{N^2}
$$

for the case $p = 2$.

5.2.4 Control of the outer interaction terms

Now we move on to the upper bounds for the terms B_1, B_2. Recall that they are defined by

$$B_a \coloneqq \frac{N-k}{N-1} \sum_{i\in[k]} \int_{\mathbb{T}^{dk}} \big(h_t^{N,k}\big)^{p-1} \nabla_i h_t^{N,k} \cdot \Big\langle K_a(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t \Big\rangle m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}),$$

where $p = 1$, 2 and $a = 1$, 2.

Control of the regular part B_2

For the term B_2, we notice that in the entropic case, we have by the Pinsker inequality

$$\Big|\Big\langle K_2(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t \Big\rangle\Big| \leqslant \|K_2\|_{L^\infty} \sqrt{2H\big(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\big)},$$

and in the L^2 case, we have

$$\Big|\Big\langle K_2(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t \Big\rangle\Big| \leqslant \|K_2\|_{L^\infty} \sqrt{D\big(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\big)}.$$

In both cases, we apply the Cauchy–Schwarz inequality

$$\begin{aligned}
\big(h_t^{N,k}\big)^{p-1} \nabla_i h_t^{N,k} \cdot \Big\langle K_a(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t \Big\rangle \\
\leqslant \frac{\varepsilon_3(N-1)}{N-k} \big(h_t^{N,k}\big)^{p-2} \big|\nabla_i h_t^{N,k}\big|^2 \\
+ \frac{(N-k)}{4\varepsilon_3(N-1)} \Big|\Big\langle K_2(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t \Big\rangle\Big|^2.
\end{aligned}$$

Integrating against the measure $m_t^{\otimes k}$ and summing over $i \in [k]$, we get

$$\begin{aligned}
B_2 &\leqslant \varepsilon_3 \mathcal{E}_p^k + \frac{\|K_2\|_{L^\infty}^2 (N-k)^2 k}{4\varepsilon_3 (N-1)^2} \\
&\qquad \times \begin{cases} \int_{\mathbb{T}^{kd}} 2H\big(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\big) m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) & \text{when } p = 1 \\ \int_{\mathbb{T}^{kd}} D\big(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\big) m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) & \text{when } p = 2 \end{cases} \\
&= \varepsilon_3 \mathcal{E}_p^k + \frac{\|K_2\|_{L^\infty}^2 (N-k)^2 k}{2p\varepsilon_3 (N-1)^2} \big(\mathcal{D}_p^{k+1} - \mathcal{D}_p^k\big).
\end{aligned}$$

The last equality is a "towering" property of relative entropy and χ^2 distance, which can be verified directly from the definition of conditional density.

Control of the singular part B_1

By the same Cauchy–Schwarz inequality as in the previous step, the term B_1 satisfies

$$\begin{aligned}
B_1 \leqslant \varepsilon_4 \mathcal{E}_p^k + \frac{(N-k)^2 k}{4\varepsilon_4 (N-1)^2} \\
\times \int_{\mathbb{T}^{kd}} \big(h_t^{N,k}\big)^p \Big|\Big\langle K_1(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t \Big\rangle\Big|^2 m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}).
\end{aligned}$$

In the entropic case where $p = 1$, applying the first inequality of Proposition 5.7 in Section 5.4 with $m_1 \to m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})$, $m_2 \to m_t$, we get

$$\begin{aligned}
&\Bigl|\Bigl\langle K_1(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t\Bigr\rangle\Bigr|^2 \\
&\quad \leqslant \|V\|_{L^\infty}^2(1+\varepsilon_5) I\bigl(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\bigr) \\
&\qquad + 2\|V\|_{L^\infty}^2(1+\varepsilon_5^{-1})\|\nabla \log m_t\|_{L^\infty}^2 H\bigl(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\bigr).
\end{aligned}$$

Noticing that the conditional entropy and Fisher information satisfy the towering property:

$$\begin{aligned}
\int_{\mathbb{T}^{kd}} H\bigl(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\bigr) m_t^{N,k}(\mathrm{d}\boldsymbol{x}^{[k]}) &= H_t^{k+1} - H_t^k, \\
\int_{\mathbb{T}^{kd}} I\bigl(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\bigr) m_t^{N,k}(\mathrm{d}\boldsymbol{x}^{[k]}) &= \frac{I_t^{k+1}}{k+1},
\end{aligned}$$

we integrate the equality above with respect to $m_t^{N,k}$ and obtain

$$\begin{aligned}
B_1 \leqslant \varepsilon_4 I_t^k &+ \frac{(1+\varepsilon_5)\|V\|_{L^\infty}^2(N-k)^2 k}{4\varepsilon_4(N-1)^2(k+1)} I_t^{k+1} \\
&+ \frac{(1+\varepsilon_5^{-1})\|V\|_{L^\infty}^2\|\nabla \log m_t\|_{L^\infty}^2(N-k)^2 k}{2\varepsilon_4(N-1)^2}\bigl(H_t^{k+1} - H_t^k\bigr).
\end{aligned}$$

In the L^2 case where $p = 2$, we apply the second inequality of Proposition 5.7 in Section 5.4 with $m_1 \to m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})$, $m_2 \to m_t$, and get

$$\begin{aligned}
&\Bigl|\Bigl\langle K_2(x^i - \cdot), m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]}) - m_t\Bigr\rangle\Bigr|^2 \\
&\quad \leqslant M_V(1+\varepsilon_5) E\bigl(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\bigr) \\
&\qquad + M_V(1+\varepsilon_5^{-1})\|\nabla \log m_t\|_{L^\infty}^2 D\bigl(m_t^{N,(k+1)|k}(\cdot|\boldsymbol{x}^{[k]})\big|m_t\bigr).
\end{aligned}$$

for $M_V \coloneqq \sup_{t\in[0,T]}\sup_{x\in\mathbb{T}^d}\int_{\mathbb{T}^d}|V(x-y)|^2 m_t(\mathrm{d}y)$. Noticing that the towering property holds for χ^2 distance and Dirichlet energy:

$$\begin{aligned}
\int_{\mathbb{T}^{kd}} \bigl(h_t^{N,k}\bigr)^2 D\bigl(m_{t,\boldsymbol{x}^{[k]}}^{N,(k+1)|k}\big|m_t\bigr) m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) &= D_t^{k+1} - D_t^k, \\
\int_{\mathbb{T}^{kd}} \bigl(h_t^{N,k}\bigr)^2 E\bigl(m_{t,\boldsymbol{x}^{[k]}}^{N,(k+1)|k}\big|m_t\bigr) m_t^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) &= \frac{E_t^{k+1}}{k+1},
\end{aligned}$$

we integrate against $m_t^{\otimes k}$ and get

$$\begin{aligned}
B_1 \leqslant \varepsilon_4 E_t^k &+ \frac{(1+\varepsilon_5)M_V(N-k)^2 k}{4\varepsilon_4(N-1)^2(k+1)} E_t^{k+1} \\
&+ \frac{(1+\varepsilon_5^{-1})M_V\|\nabla \log m_t\|_{L^\infty}^2(N-k)^2 k}{4\varepsilon_4(N-1)^2}\bigl(D_t^{k+1} - D_t^k\bigr).
\end{aligned}$$

5.2.5 Conclusion of the proof

By combining the upper bounds on A_1, A_2, B_1, B_2 obtained in the previous steps, we get

$$\begin{aligned}\frac{\mathrm{d}H_t^k}{\mathrm{d}t} \leqslant &-\big(1-\textstyle\sum_{n=1}^4 \varepsilon_n\big) I_t^k + \frac{(1+\varepsilon_5)\|V\|_{L^\infty}^2}{4\varepsilon_4} I_t^{k+1}\mathbb{1}_{k<N} \\ &+ CM_{V,m_t} H_t^k \\ &+ \bigg(\frac{C\|K_2\|_{L^\infty}^2}{\varepsilon_3} + \frac{(1+\varepsilon_5^{-1})\|V\|_{L^\infty}^2 \|\nabla \log m_t\|_{L^\infty}^2}{2\varepsilon_4}\bigg) k\big(H_t^{k+1} - H_t^k\big)\mathbb{1}_{k<N} \\ &+ CM_{V,m_t}\frac{k^2}{N^2} + C\bigg(\frac{\|K_2\|_{L^\infty}^2}{\varepsilon_1} + \frac{M_{V,m_t}}{\varepsilon_2}\bigg)\frac{k^2}{N^2} \times \begin{cases} k \\ 1 + k\sqrt{H_t^3} \end{cases}\end{aligned}$$

for the entropic case $p=1$, and

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}D_t^k}{\mathrm{d}t} \leqslant &-\big(1-\textstyle\sum_{n=1}^4 \varepsilon_n\big) E_t^k + \frac{(1+\varepsilon_5)M_V}{4\varepsilon_4} E_t^{k+1}\mathbb{1}_{k<N} \\ &+ C\bigg[M_{V,m_t}\bigg(1 + \frac{k^2}{N} + \frac{k^3}{\varepsilon_2 N^2}\bigg) + \frac{\|K_2\|_{L^\infty}^2 k^3}{N^2}\bigg] D_t^k \\ &+ \bigg(\frac{C\|K_2\|_{L^\infty}^2}{\varepsilon_3} + \frac{(1+\varepsilon_5^{-1})M_V \|\nabla \log m_t\|_{L^\infty}^2}{4\varepsilon_4}\bigg) k\big(D_t^{k+1} - D_t^k\big)\mathbb{1}_{k<N} \\ &+ C\bigg(\frac{\|K_2\|_{L^\infty}^2}{\varepsilon_1} + M_{V,m_t}(1+\varepsilon_2^{-1})\bigg)\frac{k^2}{N^2}\end{aligned}$$

for the L^2 case $p=2$.

Since $\|V\|_{L^\infty}^2$, M_V are respectively supposed to be smaller than 1 in Theorems 5.1 and 5.2, we can take

$$\varepsilon_4 = \begin{cases} \|V\|_{L^\infty}/2 & \text{when } p=1, \\ \sqrt{M_V}/2 & \text{when } p=2. \end{cases}$$

so that for ε_1, ε_2, ε_3, ε_5 small enough, we have

$$1 - \sum_{n=1}^4 \varepsilon_n > \frac{(1+\varepsilon_5)}{4\varepsilon_4} \cdot \begin{cases} \|V\|_{L^\infty}^2 & \text{when } p=1, \\ M_V & \text{when } p=2. \end{cases}$$

Additionally, for the second assertion of Theorem 5.1, since we have

$$\frac{r_*}{8\pi^2(1-\|V\|_{L^\infty})} \leqslant 1,$$

we can pick the ε_i, for $i \in [3]$ and $i=5$, such that

$$1 - \sum_{n=1}^4 \varepsilon_n - \frac{(1+\varepsilon_5)}{4\varepsilon_4}\|V\|_{L^\infty}^2 = 1 - \frac{2+\varepsilon_5}{2}\|V\|_{L^\infty} - \sum_{i=1}^3 \varepsilon_i \geqslant \frac{r_*}{8\pi^2}.$$

Fix these choices of ε_i for $i \in [5]$ in the respective situations.

Then, for the first assertion of Theorem 5.1, we choose the first alternative in the upper bound of $\mathrm{d}H_t^k/\mathrm{d}t$, and get

$$\frac{\mathrm{d}H_t^k}{\mathrm{d}t} \leqslant -c_1 I_t^k + c_2 I_t^{k+1}\mathbb{1}_{k<N} + M_1' H_t^k + M_2' k\big(H_t^{k+1} - H_t^k\big)\mathbb{1}_{k<N} + M_3'\frac{k^3}{N^2},$$

for $c_1 > c_2 \geqslant 0$ and some set of constants M'_i, $i \in [3]$. Applying the first case of Proposition 5.5 in Section 5.3 to the system of differential inequalities of H^k_t, I^k_t, we get an M' such that $H^k_t \leqslant M' e^{M't} k^3/N^2$. So taking $k = 3$, we get the bound on the 3-marginal's relative entropy: $H^3_t \leqslant 27M' e^{M't}/N^2$. Plugging this bound into the second alternative in the upper bound of $\mathrm{d}H^k_t/\mathrm{d}t$, we get

$$\frac{\mathrm{d}H^k_t}{\mathrm{d}t} \leqslant -c_1 I^k_t + c_2 I^{k+1}_t \mathbb{1}_{k<N} + M_1 H^k_t + M_2 k\big(H^{k+1}_t - H^k_t\big)\mathbb{1}_{k<N} + M_3 e^{M_3 t}\frac{k^2}{N^2},$$

for some other set of constants M_i, $i \in [3]$. We apply again the first case of Proposition 5.5 to obtain the desired result $H^k_t \leqslant M e^{Mt} k^2/N^2$.

For the second assertion of Theorem 5.1, we have $K_2 = 0$ and

$$\|\nabla \log m_t\|^2_{L^\infty} + \|\nabla^2 \log m_t\|_{L^\infty} \leqslant M_m e^{-\eta t}.$$

Taking the first alternative in the upper bound of $\mathrm{d}H^k_t/\mathrm{d}t$, we get

$$\begin{aligned}\frac{\mathrm{d}H^k_t}{\mathrm{d}t} &\leqslant -c_1 I^k_t + c_2 I^{k+1}_t \mathbb{1}_{k<N} \\ &\qquad + C M_m e^{-\eta t} H^k_t + C(1+\varepsilon_5^{-1}) M_m e^{-\eta t} k\big(H^{k+1}_t - H^k_t\big)\mathbb{1}_{k<N} \\ &\qquad\qquad + C(1+\varepsilon_2^{-1}) M_m e^{-\eta t}\frac{k^3}{N^2}.\end{aligned}$$

Notice that by our choice of constants, we have

$$c_1 - c_2 \geqslant \frac{r_*}{8\pi^2}.$$

On the other hand, according to [12, Proposition 5.7.5], the uniform measure 1 on $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ verifies a log-Sobolev inequality:

$$\forall m \in \mathcal{P}(\mathbb{T}) \text{ regular enough}, \qquad 8\pi^2 H(m|1) \leqslant I(m|1),$$

and the inequality with the same $8\pi^2$ constant for the uniform measure on $\mathbb{T}^d$ by tensorization property. By the gradient bound $\|\nabla \log m_t\|^2_{L^\infty} \leqslant M_m e^{-\eta t}$, we can control the oscillation of $\log m_t$:

$$\sup_{\mathbb{T}^d} \log m_t - \inf_{\mathbb{T}^d} \log m_t \leqslant \frac{M_m\sqrt{d}}{2} e^{-\eta t}.$$

Thus, by Holley–Stroock's perturbation result [113], the measure m_t satisfies a log-Sobolev inequality with constant

$$8\pi^2 \exp\bigg(-\frac{M_m\sqrt{d}}{2}e^{-\eta t}\bigg),$$

which implies

$$I^k_t \geqslant \frac{r_*}{c_1 - c_2} H^k_t,$$

for sufficiently large t. Let $r \in (0, r_*)$ be arbitrary. We can apply the second case of Proposition 5.5 and get

$$H^k_t \leqslant M'' e^{-rt}\frac{k^3}{N^2}.$$

We then plug the bound for H^3_t back to the second alternative in the upper bound for $\mathrm{d}H^k_t/\mathrm{d}t$ to get

$$\begin{aligned}\frac{\mathrm{d}H^k_t}{\mathrm{d}t} &\leqslant -c_1 I^k_t + c_2 I^{k+1}_t \mathbb{1}_{k<N} \\ &\qquad + CM_m e^{-\eta t} H^k_t + C(1+\varepsilon_5^{-1}) M_m e^{-\eta t} k\big(H^{k+1}_t - H^k_t\big)\mathbb{1}_{k<N} \\ &\qquad\qquad + C(1+\varepsilon_2^{-1}) M_m (1+M'') e^{-\eta t} \frac{k^2}{N^2}.\end{aligned}$$

Applying again the second case of Proposition 5.5, we obtain the desired control

$$H^k_t \leqslant M' e^{-rt} \frac{k^2}{N^2}.$$

Finally, in the L^2 case, we apply the crude bounds $k^2/N \leqslant k$, $k^3/N^2 \leqslant k$, $D^k_t \leqslant D^{k+1}_t$ in the second line of the upper bound for $\mathrm{d}D^k_t/\mathrm{d}t$, and $k\big(D^{k+1}_t - D^k_t\big) \leqslant kD^{k+1}_t$ in the third line. So we get

$$\frac{\mathrm{d}D^k_t}{\mathrm{d}t} \leqslant -c_1 E^k_t + c_2 E^{k+1}_t \mathbb{1}_{k<N} + M_2 k D^{k+1}_t \mathbb{1}_{k<N} + M_3 \frac{k^2}{N^2}$$

for some $c_1 > c_2 \geqslant 0$ and M_2, $M_3 \geqslant 0$. We conclude the proof by applying Proposition 5.6 in Section 5.3 to the system of D^k_t, E^k_t. □

5.3 ODE hierarchies

5.3.1 Entropic hierarchy

Now we move on to solving the ODE hierarchy that is "weaker" than that considered in [140]. As we have seen in the previous section, in the time-derivative of the k-th level entropy $\mathrm{d}H^k_t/\mathrm{d}t$, we allow the Fisher information of the next level, i.e. I^{k+1}_t, to appear. In this section, we show that as long as the extra term's coefficient is controlled by the heat dissipation, the hierarchy still preserves the $O(k^2/N^2)$ order globally in time. This is achieved by choosing a weighted mix of entropies at all levels $\geqslant k$ so that when we consider its time-evolution, a telescoping sequence appears and cancels all the Fisher informations.

Proposition 5.5. *Let $T \in (0,\infty]$ and let $x^k_\cdot$, $y^k_\cdot : [0,T) \to \mathbb{R}_{\geqslant 0}$ be $\mathcal{C}^1$ functions, for $k \in [N]$. Suppose that $x^{k+1}_t \geqslant x^k_t$ for all $k \in [N-1]$. Suppose that there exist integer $\beta \geqslant 2$, real numbers $c_1 > c_2 \geqslant 0$ and $C_0 \geqslant 0$, and functions M_1, M_2, $M_3 : [0,T) \to [0,\infty)$ such that for all $t \in [0,T)$ and $k \in [N]$, we have*

$$\begin{aligned}x^k_0 &\leqslant \frac{C_0 k^2}{N^2}, \\ \frac{\mathrm{d}x^k_t}{\mathrm{d}t} &\leqslant -c_1 y^k_t + c_2 y^{k+1}_t \mathbb{1}_{k<N} + M_1(t) x^k_t + M_2(t) k\big(x^{k+1}_t - x^k_t\big)\mathbb{1}_{k<N} + M_3(t)\frac{k^\beta}{N^2}.\end{aligned} \tag{5.6}$$

Then we have the following results.

1. *If M_1, M_2 are constant functions and $M_3(t) \leqslant Le^{Lt}$ for some $L \geqslant 0$, then there exists $M > 0$, depending only on β, c_1, c_2, C_0, M_1, M_2 and L, such that for all $t \in [0,T)$, we have*

$$x^k_t \leqslant M e^{Mt} \frac{k^\beta}{N^2}.$$

2. If $T = \infty$, the functions M_1, M_2, M_3 are non-increasing and satisfy

$$M_i(t) \leqslant Le^{-\eta t}$$

for all $t \in [0, \infty)$ and all $i \in [3]$, for some $L \geqslant 0$, $\eta > 0$ and if $y_t^k \geqslant \rho x_t^k$ for all $t \in [t_*, \infty)$ for some $\rho > 0$ and some $t_* \geqslant 0$, then for all $r \in \big(0, \rho(c_1 - c_2)\big)$, there exists $M' \geqslant 0$, depending only on r, η, β, c_1, c_2, C_0, L, ρ and t_*, such that for all $t \in [0, \infty)$, we have

$$x_t^k \leqslant M' e^{-\min(r,\eta)t} \frac{k^\beta}{N^2}.$$

Proof. We prove the proposition by considering the two cases at the same time. Notice that the relation

$$y_t^k \geqslant \rho x_t^k$$

trivially holds for $\rho = 0$. We set $t_* = \infty$ in the first case. Allowing ρ to be a function of time, we simply set $\rho(\cdot) = 0$ in the first situation and in the second situation on the interval $[0, t_*]$ for the rest of the proof. So formally we can write

$$\rho(t) = \rho \mathbb{1}_{t \geqslant t_*}.$$

To avoid confusion we will always write $\rho(\cdot)$ for the time-dependent function and ρ for the constant.

Step 1: Reduction to $M_1 = 0$. We first notice that, by defining the new variables

$$x_t'^k = x_t^k \exp\left(-\int_0^t M_1(s)\,\mathrm{d}s\right),\ y_t'^k = y_t^k \exp\left(-\int_0^t M_1(s)\,\mathrm{d}s\right),$$

we can reduce to the case where $M_1 = 0$ upon redefining M_3 (and therefore L in the second case, but not η). This transform does not change the relations

$$x_t^{k+1} \geqslant x_t^k, \qquad y_t^k \geqslant \rho x_t^k$$

and the initial values of x^k, so we can suppose $M_1 = 0$ without loss of generality.

Step 2: Reduction to $k \leqslant N/2$. Second, by taking $k = N$ in the hierarchy (5.6), we find

$$\frac{\mathrm{d}x_t^N}{\mathrm{d}t} \leqslant -\rho(t) x_t^N + M_3(t) N^{\beta-2}$$

and thus the a priori bound follows:

$$x_t^N \leqslant \left(C_0 e^{-\int_0^t \rho} + \int_0^t e^{-\int_s^t \rho} M_3(s)\,\mathrm{d}s\right) N^{\beta-2} =: M_t^N N^{\beta-2} \tag{5.7}$$

In the second case where $\rho(\cdot)$ is eventually constant: $\rho(\cdot) = \rho > 0$, the quantity M_t^N is exponentially decreasing in t with rate $\min(\rho, \eta)$. By the monotonicity of $k \mapsto x_t^k$, we get that for all $k > N/2$,

$$x_t^k \leqslant x_t^N \leqslant M_t^N N^{\beta-2} < 2^\beta M_t^N \frac{k^\beta}{N^2}.$$

So it only remains to establish the upper bound of x_t^k for $k \leqslant N/2$.

Step 3: New hierarchy. Let α be an arbitrary real number $\geqslant \beta+3$. Recall that in the second case, $r \in \big(0, \rho(c_1-c_2)\big)$ and in the first case we simply set $r=0$ and adopt the convention $0/0=0$. Let

$$i_0 \coloneqq \max\left(1, \inf\left\{ i>0 : \frac{i^\alpha}{(i+1)^\alpha} \geqslant \frac{c_2+r/\rho}{c_1}\right\}\right).$$

The number i_0 is always well defined, as $\lim_{i\to\infty} i^\alpha/(i+1)^\alpha = 1 > (c_2+r/\rho)/c_1$. Thus, for any $i \geqslant i_0$, we have

$$\frac{c_1}{(i+1)^\alpha} \geqslant \frac{c_2}{i^\alpha} + \frac{r}{\rho i^\alpha}.$$

Define, for $k\in[N]$ and $t\geqslant 0$, the following new variable:

$$z_t^k \coloneqq \sum_{i=k}^N \frac{x_t^i}{(i-k+i_0)^\alpha}.$$

By summing up the ODE hierarchy (5.6) (with $M_1=0$), we find

$$\begin{aligned}\frac{\mathrm{d}z_t^k}{\mathrm{d}t} \leqslant -\sum_{i=k}^N \frac{c_1 y_t^i}{(i-k+i_0)^\alpha} + \sum_{i=k}^{N-1} \frac{c_2 y_t^{i+1}}{(i-k+i_0)^\alpha} \\ + \frac{M_3(t)}{N^2}\sum_{i=k}^N \frac{i^\beta}{(i-k+i_0)^\alpha} + M_2(t)\sum_{i=k}^{N-1}\frac{i}{(i-k+i_0)^\alpha}\big(x_t^{i+1}-x_t^i\big).\end{aligned}$$

The sum of the first two terms satisfy

$$\begin{aligned}&-\sum_{i=k}^N \frac{c_1 y_t^i}{(i-k+i_0)^\alpha} + \sum_{i=k}^{N-1} \frac{c_2 y_t^{i+1}}{(i-k+i_0)^\alpha} \\ &\quad= -\frac{c_1 y_t^k}{i_0^\alpha} + \sum_{i=k}^N\left(-\frac{c_1}{(i+1-k+i_0)^\alpha} + \frac{c_2}{(i-k+i_0)^\alpha}\right) y_t^i \\ &\quad\leqslant -\sum_{i=k}^N \frac{r\rho(t) y_t^i}{\rho(i-k+i_0)^\alpha} \leqslant -\sum_{i=k}^N \frac{r x_t^i}{(i-k+i_0)^\alpha} = -r z_t^k \mathbb{1}_{t\geqslant t_*},\end{aligned}$$

thanks to our choice of i_0. For the third term, we find

$$\begin{aligned}\sum_{i=k}^N \frac{i^\beta}{(i-k+i_0)^\alpha} \leqslant C_\beta \sum_{i=k}^N \frac{(i-k)^\beta + k^\beta}{(i-k+1)^\alpha} \leqslant C_\beta \sum_{i=1}^\infty \frac{(i-1)^\beta}{i^\alpha} + C_\beta k^\beta \sum_{i=1}^\infty \frac{1}{i^\alpha} \\ \leqslant C_{\alpha,\beta} k^\beta, \quad (5.8)\end{aligned}$$

where $C_\beta > 0$ (resp. $C_{\alpha,\beta}>0$) depends only on β (resp. α and β). In the following, we allow these constants to change from line to line. For the last term, we perform the summation by parts:

$$\begin{aligned}&\sum_{i=k}^{N-1} \frac{i}{(i-k+i_0)^\alpha}\big(x_t^{i+1}-x_t^i\big) \\ &\quad= -\frac{k}{i_0^\alpha} x_t^k + \frac{N}{(N-k+i_0)^\alpha} x_t^N + \sum_{i=k}^{N-1}\left(\frac{i}{(i-k+i_0)^\alpha} - \frac{(i+1)}{(i+1-k+i_0)^\alpha}\right) x_t^{i+1}.\end{aligned}$$

The coefficient in the last summation satisfies

$$\begin{aligned}
&\frac{i}{(i-k+i_0)^\alpha} - \frac{(i+1)}{(i+1-k+i_0)^\alpha} \\
&\quad = \left(\frac{1}{(i-k+i_0)^{\alpha-1}} - \frac{1}{(i+1-k+i_0)^{\alpha-1}}\right) \\
&\qquad + (k-i_0)\left(\frac{1}{(i-k+i_0)^\alpha} - \frac{1}{(i+1-k+i_0)^\alpha}\right) \\
&\quad \leqslant \frac{\alpha-1}{(i-k+i_0)^\alpha} + k\left(\frac{1}{(i-k+i_0)^\alpha} - \frac{1}{(i+1-k+i_0)^\alpha}\right),
\end{aligned}$$

where the last inequality is due to $j^{-\alpha+1} - (j+1)^{-\alpha+1} \leqslant (\alpha-1)j^{-\alpha}$ for $\alpha > 1$ and $j > 0$. Thus, we have

$$\begin{aligned}
&\sum_{i=k}^{N-1} \frac{i}{(i-k+i_0)^\alpha}\left(x_t^{i+1} - x_t^i\right) \\
&\quad \leqslant -\frac{k}{i_0^\alpha} x_t^k + \frac{N}{(N-k+i_0)^\alpha} x_t^N + (\alpha-1)\sum_{i=k}^{N-1} \frac{x_t^{i+1}}{(i-k+i_0)^\alpha} \\
&\qquad + k\sum_{i=k}^{N-1}\left(\frac{1}{(i-k+i_0)^\alpha} - \frac{1}{(i+1-k+i_0)^\alpha}\right) x_t^{i+1}
\end{aligned}$$

The difference between z_t^{k+1} and z_t^k reads

$$z_t^{k+1} - z_t^k = \sum_{i=k}^{N-1}\left(\frac{1}{(i-k+i_0)^\alpha} - \frac{1}{(i+1-k+i_0)^\alpha}\right) x_t^{i+1} - \frac{x_t^k}{i_0^\alpha}.$$

Then, rewriting in terms of z_t^k and z_t^{k+1}, we find that, for $k \in [N-1]$, the last summation satisfies

$$\begin{aligned}
&\sum_{i=k}^{N-1} \frac{i}{(i-k+i_0)^\alpha}\left(x_t^{i+1} - x_t^i\right) \\
&\quad \leqslant \sum_{i=k}^{N-1} \frac{\alpha-1}{(i-k+i_0)^\alpha} x_t^{i+1} + k\left(z_t^{k+1} - z_t^k\right) + \frac{N}{(N-k+i_0)^\alpha} x_t^N \\
&\quad \leqslant \frac{(\alpha-1)c_1}{c_2} \sum_{i=k+1}^{N} \frac{x_t^i}{(i-k+i_0)^\alpha} + k\left(z_t^{k+1} - z_t^k\right) + \frac{N}{(N-k+i_0)^\alpha} x_t^N \\
&\quad = \frac{(\alpha-1)c_1}{c_2} z_t^k + k\left(z_t^{k+1} - z_t^k\right) + \frac{N}{(N-k+i_0)^\alpha} x_t^N.
\end{aligned}$$

Then for $k \leqslant N/2$, we have

$$\begin{aligned}
&\sum_{i=k}^{N-1} \frac{i}{(i-k+i_0)^\alpha}\big(x_t^{i+1} - x_t^i\big) \\
&\quad\leqslant \frac{(\alpha-1)c_1}{c_2} z_t^k + k\big(z_t^{k+1} - z_t^k\big) + \frac{N}{(N/2)^\alpha} x_t^N \\
&\quad\leqslant \frac{(\alpha-1)c_1}{c_2} z_t^k + k\big(z_t^{k+1} - z_t^k\big) + \frac{2^\alpha}{N^{\alpha-1}} M_t^N N^{\beta-2} \\
&\quad\leqslant \frac{(\alpha-1)c_1}{c_2} z_t^k + k\big(z_t^{k+1} - z_t^k\big) + \frac{2^\alpha M_t^N}{N^2},
\end{aligned}$$

where the last inequality is due to $\alpha \geqslant \beta + 3$. Combining the upper bounds for all the terms, we get, for $k \leqslant N/2$,

$$\begin{aligned}
\frac{\mathrm{d}z_t^k}{\mathrm{d}t} \leqslant -r z_t^k \mathbb{1}_{t \geqslant t_*} + \frac{(\alpha-1)c_1 M_2(t)}{c_2} z_t^k + M_2(t) k \big(z_t^{k+1} - z_t^k\big) \\
+ C_{\alpha,\beta} M_3(t) \frac{k^\beta}{N^2} + \frac{2^\alpha M_t^N M_2(t)}{N^2}, \quad (5.9)
\end{aligned}$$

For $k = \bar{k} \coloneqq \lfloor N/2 \rfloor + 1$, we have by the a priori bound (5.7),

$$z_t^{\bar{k}} = \sum_{i=\bar{k}}^{N} \frac{x_t^i}{(i-\bar{k}+i_0)^\alpha} \leqslant x_t^N \sum_{i=k}^{N} \frac{1}{(i-\bar{k}+i_0)^\alpha} \leqslant C_\alpha M_t^N N^{\beta-2}.$$

According to the computations in (5.8), the initial values of z_0^k, for $k \leqslant N/2$, satisfy

$$z_0^k \leqslant C_\alpha C_0 \frac{k^2}{N^2} =: C_0' \frac{k^2}{N^2}.$$

So the new hierarchy in terms of z_t^k is derived.

At this point, we can already apply the Grönwall iteration method of Lacker [140] and, in the time-uniform case, of Lacker and Le Flem [142], to solve the system of differential inequalities (5.9). However, we take a much simpler approach here based on the following observation. If the variable k in (5.9) is no longer discrete but continuous, then the term $M_2(t)k\big(z_t^{k+1} - z_t^k\big)$ becomes the transport term

$$M_2(t) k \frac{\partial z_t^{k+1}}{\partial k},$$

and z_t^k becomes a subsolution to a transport equation

$$\frac{\partial z_t^k}{\partial t} \leqslant -r z_t^k \mathbb{1}_{t \geqslant t_*} + M_2(t) k \frac{\partial z_t^k}{\partial k} + \text{source terms}.$$

Since the transport equation verifies a comparison principle, it suffices to construct a supersolution to the equation that dominates z_t^k on the parabolic boundary, in order to obtain an upper bound for z_t^k in the continuous case. The crucial observation here, which we prove in Proposition 5.11 in Section 5.4.3, is that the comparison still holds for the discretization scheme (5.9). So in the following we construct supersolutions for the system of differential inequalities in the two cases of the proposition.

Step 4.1: Global-in-time estimates. In the first case, we can control M_t^N defined in (5.7) by

$$M_t^N \leqslant C_0 + e^{Lt} - 1.$$

Thus, by the last step,

$$z_t^{\bar{k}} \leqslant C_\alpha (C_0 + e^{Lt} - 1) N^{\beta-2}.$$

where $\bar{k} = \lfloor N/2 \rfloor + 1$ as we recall. Now we set, for $k \leqslant N/2$,

$$w_t^k = M e^{Mt} \frac{k^\beta}{N^2}$$

for some M to be determined. For M large enough, we have the domination

$$w_t^k \geqslant z_t^k$$

on the parabolic boundary

$$\{(t,k) \in [0,\infty) \times [N] : t = 0 \text{ or } k = \bar{k}\}.$$

In the interior, w_t^k is an upper solution for (5.9) if and only if

$$M^2 e^{Mt} \frac{k^\beta}{N^2} \geqslant \frac{(\alpha-1)c_1 M_2}{c_2} M e^{Mt} \frac{k^\beta}{N^2} + M_2 \frac{k\big((k+1)^\beta - k^\beta\big)}{N^2} + C_{\alpha,\beta} \frac{k^\beta}{N^2} + 2^\alpha M_2 \frac{C_0 + e^{Lt} - 1}{N^2}.$$

Noting that $(k+1)^\beta - k^\beta \leqslant \beta(k+1)^{\beta-1} \leqslant 2^{\beta-1}\beta k^{\beta-1}$, we can let the inequality hold by taking an M large enough. We conclude in this case by applying the comparison principle of Proposition 5.11 to $w_t^k - z_t^k$.

Step 4.2: Exponentially decaying estimate. In this case, the a priori bound M_t^N verifies, for some $M'' > 0$,

$$M_t^N \leqslant M'' e^{-\min(r,\eta)t}.$$

We set, for $k \leqslant N/2$,

$$w_t^k = M'(t) \frac{k^\beta}{N^2}$$

for some $M' : [0,\infty) \to [0,\infty)$ to be determined. The domination $w_t^k \geqslant z_t^k$ on the boundary is satisfied if

$$M'(0) \geqslant C_0'$$
$$M'(t) \geqslant C_\alpha M_t^N,$$

In the interior, w_t^k is an upper solution for (5.9) if and only if

$$\frac{\mathrm{d}M'(t)}{\mathrm{d}t} \geqslant -r \mathbb{1}_{t \geqslant t_*} M'(t) + \frac{(\alpha-1)c_1 M_2(t)}{c_2} M'(t) + M_2(t) \frac{k\big((k+1)^\beta - k^\beta\big)}{k^\beta} M'(t) + C_{\alpha,\beta} M_3(t) + \frac{2^\alpha M_t^N M_2(t)}{k^\beta}.$$

Note that the source terms on the second line can be bounded by $L''e^{-\eta t}$ for some $L'' > 0$. Set

$$\rho'(t) = r\mathbb{1}_{t\geqslant t_*} - \left(\frac{(\alpha-1)c_1}{c_2} + 2^{\beta-1}\beta\right)M_2(t)$$

and

$$M'(t) = M_0' e^{-\int_0^t \rho'} + \int_0^t e^{-\int_s^t \rho'} L'' e^{-\eta s}\,\mathrm{d}s.$$

We find that all conditions are satisfied for an M_0' sufficiently large. We fix such M_0' and apply again Proposition 5.11 to $w_t^k - z_t^k$ to conclude. □

5.3.2 L^2 hierarchy

For the ODE system obtained from the L^2 hierarchy, we only show that the $O(1/N^2)$-order bound holds until some finite time. We note that similar hierarchies have appeared recently in [30, 29].

Proposition 5.6. *Let $T > 0$ and let $x_\cdot^k$, $y_\cdot^k : [0,T] \to \mathbb{R}_{\geqslant 0}$ be $\mathcal{C}^1$ functions, for $k \in [N]$. Suppose that there exist real numbers $c_1 > c_2 \geqslant 0$, and C_0, M_2, $M_3 \geqslant 0$ such that for all $t \in [0,T]$ and $k \in [N]$, we have*

$$\begin{aligned} x_0^k &\leqslant \frac{C_0 k^2}{N^2}, \\ \frac{\mathrm{d}x_t^k}{\mathrm{d}t} &\leqslant -c_1 y_t^k + c_2 y_t^{k+1}\mathbb{1}_{k<N} + M_2 k x_t^{k+1}\mathbb{1}_{k<N} + M_3\frac{k^2}{N^2}. \end{aligned}$$

Then, there exist T_*, $M > 0$, depending only on β, c_1, c_2, C_0, M_2, M_3, such that for all $t \in [0, T_* \wedge T)$, we have

$$x_t^k \leqslant \frac{Me^{Mk}}{(T_* - t)^3 N^2}.$$

Proof. For $t \in [0,T]$ and $r \in [c_2/c_1, 1]$, we define the generating function (or the Laplace transform) associated to x_t^k:

$$F(t,r) = \sum_{k=1}^N r^k x_t^k.$$

Then, taking the time-derivative of $F(t,r)$, we get

$$\begin{aligned} \frac{\partial F(t,r)}{\partial t} &\leqslant -c_1 \sum_{k=1}^N r^k y_t^k + c_2 \sum_{k=1}^{N-1} r^k y_t^{k+1} + M_2 \sum_{k=1}^{N-1} k r^k x_t^{k+1} + \frac{M_3}{N^2}\sum_{k=1}^N k^2 r^k \\ &\leqslant -c_1 r y_t^1 + \sum_{k=2}^N (c_2 - c_1 r) r^{k-1} y_t^{k+1} + M_2 \sum_{k=1}^{N-1} k r^k x_t^{k+1} + \frac{M_3}{N^2}\sum_{k=1}^N k^2 r^k \\ &\leqslant M_2 \sum_{k=1}^{N-1} k r^k x_t^{k+1} + \frac{M_3}{N^2}\sum_{k=1}^N k^2 r^k. \end{aligned}$$

Notice that, by taking partial derivatives in r, we get

$$\frac{\partial F(t,r)}{\partial r} = \sum_{k=0}^{N-1} (k+1) r^k x_t^{k+1},$$
$$\frac{\partial^2}{\partial r^2}\left(\frac{1}{1-r}\right) = \sum_{k=0}^{\infty} (k+2)(k+1) r^k.$$

Thus, we find

$$\frac{\partial F(t,r)}{\partial t} \leqslant M_2 \frac{\partial F(t,r)}{\partial r} + \frac{2M_3}{N^2(1-r)^3}.$$

The initial condition of F satisfies

$$F(0,r) = \sum_{k=1}^{N} r^k x_0^k \leqslant \frac{C_0}{N^2} \sum_{k=1}^{N} r^k k^2 \leqslant \frac{2C_0}{N^2(1-r)^3}.$$

Let

$$T_* = \frac{1}{M_2}\left(1 - \frac{c_2}{c_1}\right)$$

and for $t < T_* \wedge T$, let $(r_s)_{s\in[0,t]}$ be the characteristic line:

$$r_s = \frac{c_2}{c_1} + M_2(t-s).$$

We then have $r_0 \leqslant c_2/c_1 + M_2 t$. Integrating along this line, we get

$$\begin{aligned} F(t,r_t) &\leqslant F(0,r_0) + \frac{2M_3}{N^2}\int_0^t \frac{\mathrm{d}s}{(1-r_s)^3} \\ &\leqslant \frac{2C_0}{N^2(1-r_0)^3} + \frac{2M_3}{M_2 N^2}\int_{r_t}^{r_0} \frac{\mathrm{d}r}{(1-r)^3} \\ &\leqslant \left(\frac{2C_0}{(1-r_0)^3} + \frac{M_3}{M_2(1-r_0)^2}\right)\frac{1}{N^2}. \end{aligned}$$

Thus we get

$$x_t^k \leqslant r_t^{-k} F(t,r_t) \leqslant \left(\frac{c_1}{c_2}\right)^k \left(\frac{2C_0}{\left(1 - M_2 t - \frac{c_2}{c_1}\right)^3} + \frac{M_3}{M_2\left(1 - M_2 t - \frac{c_2}{c_1}\right)^2}\right)\frac{1}{N^2}. \qquad \square$$

5.4 Other technical results

5.4.1 Transport inequality for $W^{-1,\infty}$ kernels

One key ingredient of the entropic hierarchy of Lacker [140] is to control the outer interaction terms by the relative entropy through the Pinsker or Talagrand's transport inequality. In our situation, the interaction kernel is more singular, and we are no longer able to control the difference by the mere relative entropy. It turns out that the additional quantity to consider is the relative Fisher information.[5] We also include the inequality for the L^2 hierarchy here, as the two inequalities share the same form.

[5] It has been communicated to the author that Lacker has also obtained the inequality independently.

Proposition 5.7. *For all $K = \nabla \cdot V$ with $V \in L^\infty(\mathbb{T}^d; \mathbb{R}^d \times \mathbb{R}^d)$ and all regular enough measures m_1, $m_2 \in \mathcal{P}(\mathbb{T}^d)$, we have*

$$|\langle K, m_1 - m_2\rangle| \leqslant \|V\|_{L^\infty}\Big(\sqrt{I(m_1|m_2)} + \|\nabla \log m_2\|_{L^\infty}\sqrt{2H(m_1|m_2)}\Big),$$
$$|\langle K, m_1 - m_2\rangle| \leqslant \|V\|_{L^2(m_2)}\Big(\sqrt{E(m_1|m_2)} + \|\nabla \log m_2\|_{L^\infty}\sqrt{D(m_1|m_2)}\Big).$$

Proof. For the first inequality, we have

$$\begin{aligned}
&|\langle K, m_1 - m_2\rangle| \\
&\quad = |\langle V, \nabla m_1 - \nabla m_2\rangle| \\
&\quad \leqslant \int_{\mathbb{T}^d} |V| \left|\frac{\nabla m_1}{m_1} - \frac{\nabla m_2}{m_2}\right| \mathrm{d}m_1 + \int_{\mathbb{T}^d} \frac{|\nabla m_2|}{m_2} |V| \,\mathrm{d}|m_1 - m_2| \\
&\quad \leqslant \|V\|_{L^\infty}\left(\int_{\mathbb{T}^d}\left|\nabla \log \frac{m_1}{m_2}\right|^2 \mathrm{d}m_1\right)^{1/2} + \|\nabla \log m_2\|_{L^\infty}\|V\|_{L^\infty}\|m_1 - m_2\|_{L^1} \\
&\quad \leqslant \|V\|_{L^\infty}\Big(\sqrt{I(m_1|m_2)} + \|\nabla \log m_2\|_{L^\infty}\sqrt{2H(m_1|m_2)}\Big).
\end{aligned}$$

For the second inequality, we set $h = m_1/m_2$ and find

$$\begin{aligned}
&|\langle K_1, m_1 - m_2\rangle| \\
&\quad = \left|\int_{\mathbb{T}^d} K(h-1)\,\mathrm{d}m_2\right| \\
&\quad \leqslant \left|\int_{\mathbb{T}^d} V\nabla h \,\mathrm{d}m_2\right| + \left|\int_{\mathbb{T}^d} V(h-1)\nabla \log m_2 \,\mathrm{d}m_2\right| \\
&\quad \leqslant \|V\|_{L^2(m_2)}\big(\|\nabla h\|_{L^2(m_2)} + \|\nabla \log m_t\|_{L^\infty}\|h-1\|_{L^2(m_2)}\big).
\end{aligned}$$

□

5.4.2 Improved Jabin–Wang lemma

In the following we state a slight improvement to [124, Theorem 4], in the sense that we get the correct asymptotic behavior when the cumulant "test function" (ϕ as denoted in their work) tends to zero. This behavior is not needed for their global approach but is necessary for the inner interaction bound in our local approach. For simplicity, we denote the universal constant of Jabin–Wang by

$$C_{\mathrm{JW}} \coloneqq 1600^2 + 36e^4.$$

Theorem 5.8 (Alternative version of [124, Theorem 4])**.** *Let $\phi \in L^\infty(\mathbb{T}^d \times \mathbb{T}^d; \mathbb{R})$ and $m \in \mathcal{P}(\mathbb{T}^d)$ be such that $\int_{\mathbb{T}^d} \phi(x,y)m(\mathrm{d}y) = \int_{\mathbb{T}^d} \phi(y,x)m(\mathrm{d}y) = 0$ and $\phi(x,x) = 0$ for all $x \in \mathbb{T}^d$. Denote $\gamma = C_{\mathrm{JW}}\|\phi\|^2_{L^\infty}$. If $\gamma \in \big[0, \frac{1}{2}\big]$, then for all integer $k \geqslant 1$, we have*

$$\log \int_{\mathbb{T}^{kd}} \exp\left(\frac{1}{k}\sum_{i,j\in[k]} \phi(x^i, x^j)\right) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \leqslant 6\gamma.$$

The proof will depend on two combinatorical estimates in [124], which we state here for the readers' convenience.

Proposition 5.9 ([124, Propositions 4 and 5])**.** *Under the assumptions of Theorem 5.8, for all integer $r \geqslant 1$, we have*

$$\frac{1}{(2r)!}\int_{\mathbb{T}^{kd}}\left|\frac{1}{k}\sum_{i,j\in[k]}\phi(x^i,x^j)\right|^{2r} m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \leqslant \begin{cases} (6e^2\|\phi\|_{L^\infty})^{2r} & \text{if } 4r > k, \\ (1600\|\phi\|_{L^\infty})^{2r} & \text{if } 4 \leqslant 4r \leqslant k, \end{cases}$$

Proof of Theorem 5.8. Let $a \neq 0$. We have the elementary inequality

$$\begin{aligned} e^a - a - 1 &= \sum_{r=2}^{\infty} \frac{a^r}{r!} \leqslant \sum_{r=2}^{\infty} \frac{|a|^r}{r!} \\ &= \sum_{r=1}^{\infty} \frac{|a|^{2r}}{(2r)!} + \sum_{r=1}^{\infty} \frac{|a|^{2r+1}}{(2r+1)!} \\ &\leqslant \sum_{r=1}^{\infty} \frac{|a|^{2r}}{(2r)!} + \sum_{r=1}^{\infty} \frac{|a|^{2r+1}}{2(2r+1)!}\left(\frac{|a|}{2r+2} + \frac{2r+2}{|a|}\right) \\ &\leqslant 3\sum_{r=1}^{\infty} \frac{|a|^{2r}}{(2r)!}. \end{aligned}$$

The inequality $e^a - a - 1 \leqslant 3\sum_{r=1}^{\infty} \frac{|a|^{2r}}{(2r)!}$ holds true for $a = 0$ as well. Taking $a = \frac{1}{k}\sum_{i,j\in[k]} \phi(x^i, x^j)$ in the inequality above and integrating with $m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]})$, we get

$$\begin{aligned} &\int_{\mathbb{T}^{kd}} \exp\left(\frac{1}{k}\sum_{i,j\in[k]} \phi(x^i, x^j)\right) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\quad\leqslant 1 + \frac{1}{k}\sum_{i,j\in[k]} \int_{\mathbb{T}^{kd}} \phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\qquad + 3\sum_{r=1}^{\infty} \frac{1}{(2r)!}\int_{\mathbb{T}^{kd}} \left|\frac{1}{k}\sum_{i,j\in[k]} \phi(x^i, x^j)\right|^{2r} m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}). \end{aligned}$$

The second term on the right hand side vanishes, as by assumption, for $i \neq j$, we have $\int_{\mathbb{T}^{kd}} \phi(x^i, x^j) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) = 0$, and for $i = j$, we have $\phi(x^i, x^i) = 0$. Thus, using the counting result of Proposition 5.9, we get

$$\begin{aligned} &\int_{\mathbb{T}^{kd}} \exp\left(\frac{1}{k}\sum_{i,j\in[k]} \phi(x^i, x^j)\right) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \\ &\quad\leqslant 1 + 3\sum_{r=1}^{\lfloor k/4\rfloor} (1600\|\phi\|_{L^\infty})^{2r} + 3\sum_{r=\lfloor k/4\rfloor+1}^{\infty} (6e^2\|\phi\|_{L^\infty})^{2r} = 1 + \frac{3\gamma}{1-\gamma} \end{aligned}$$

We conclude by noting that $\log\bigl(1 + \frac{3\gamma}{1-\gamma}\bigr) \leqslant \frac{3\gamma}{1-\gamma} \leqslant 6\gamma$ for $\gamma \in \bigl[0, \frac{1}{2}\bigr]$. $\square$

Then, taking a rescaling of ϕ, we get the following.

Corollary 5.10. *Suppose that the function $\phi \in L^\infty(\mathbb{T}^d \times \mathbb{T}^d; \mathbb{R})$ and the measure $m \in \mathcal{P}(\mathbb{R}^d)$ satisfy $\int_{\mathbb{T}^d} \phi(x, y)m(\mathrm{d}y) = \int_{\mathbb{T}^d} \phi(y, x)m(\mathrm{d}y) = 0$ and $\phi(x, x) = 0$ for all $x \in \mathbb{T}^d$. Then, for all integer $N \geqslant 2$ and $k \in [N]$, we have*

$$\log \int_{\mathbb{T}^{kd}} \exp\left(\frac{1}{N}\sum_{i,j\in[k]} \phi(x^i, x^j)\right) m^{\otimes k}(\mathrm{d}\boldsymbol{x}^{[k]}) \leqslant 6C_{\mathrm{JW}}\|\phi\|_{L^\infty}^2 \frac{k^2}{N^2},$$

given that $C_{\mathrm{JW}}\|\phi\|_{L^\infty}^2 \leqslant 1/2$.

5.4.3 Maximum principle

We show a maximum principle for a system of ODE by standard method.

Proposition 5.11. *Let $T > 0$ and let $\boldsymbol{x} : [0,T] \to \mathbb{R}^N$ be a $\mathcal{C}^1$ continuous function. Suppose that the each component of the initial value $\boldsymbol{x}(0)$ is non-negative, i.e., $x^i(0) \geqslant 0$ for all $i \in [N]$. Suppose that it satisfies*

$$\forall t \in [0,T],\ \forall i \in [N], \qquad \frac{\mathrm{d}x^i(t)}{\mathrm{d}t} \geqslant \sum_{j \in [N]} A^i_j(t) x^j(t)$$

for some continuous matrix-valued $A : [0,T] \to \mathbb{R}^{d\times d}$ whose off-diagonal elements are non-negative, i.e., $A^i_j(t) \geqslant 0$ for all i, $j \in [N]$ such that $i \neq j$. Then, for all $t \in [0,T]$ and all $i \in [N]$, we have $x^i(t) \geqslant 0$.

Proof of Proposition 5.11. Denote

$$\|A\| = \sup_{i,j\in[N]} \sup_{t\in[0,T]} |A^i_j(t)|.$$

Let $\varepsilon > 0$. Then the new function $\boldsymbol{x}_\varepsilon : [0,T] \to \mathbb{R}^N$, defined componentwise

$$x^i_\varepsilon(t) \coloneqq x^i(t) + \varepsilon t,$$

verifies

$$\frac{\mathrm{d}x^i_\varepsilon(t)}{\mathrm{d}t} \geqslant \frac{\varepsilon}{2} + \sum_{j\in[N]} A^i_j x^j_\varepsilon(t)$$

for all $t \leqslant T \wedge (2N\|A\|)^{-1} =: T_1$. Suppose that one component of $\boldsymbol{x}_\varepsilon$ becomes negative on $[0,T_1]$. Then the following time is well defined:

$$\tau \coloneqq \inf\{t \in [0,T_1] : \exists i \in [N],\ x^i_\varepsilon(t) \leqslant 0\}.$$

Let $\iota \in [N]$ be one index such that $x^\iota_\varepsilon(\tau) \leqslant 0$, i.e., the above condition is met. By the continuity, we must have $x^\iota(\tau) = 0$ and $x^j_\varepsilon(\tau) \geqslant 0$ for all $j \neq \iota$. Then,

$$\frac{\mathrm{d}x^\iota_\varepsilon(\tau)}{\mathrm{d}t} \geqslant \frac{\varepsilon}{2} + A^\iota_\iota(\tau)x^\iota_\varepsilon(\tau) + \sum_{j\in[N]\setminus\{\iota\}} A^\iota_j(\tau)x^j_\varepsilon(\tau) \geqslant \frac{\varepsilon}{2} > 0,$$

which is a contradiction. So $x^i_\varepsilon(t) \geqslant 0$ for all $t \in [0,T_1]$ and $i \in [N]$. Taking $\varepsilon \to 0$, we get the positivity for $\boldsymbol{x}$ on $[0,T_1]$. Reiterating the proof if necessary, we get the positivity on the whole interval $[0,T]$, which is the claim of the proposition. ☐

Part III

Alternative mean field dynamics

Chapter 6

Entropic fictitious play for mean field optimization problem

Abstract. We study two-layer neural networks in the mean field limit, where the number of neurons tends to infinity. In this regime, the optimization over the neuron parameters becomes the optimization over the probability measures, and by adding an entropic regularizer, the minimizer of the problem is identified as a fixed point. We propose a novel training algorithm named entropic fictitious play, inspired by the classical fictitious play in game theory for learning Nash equilibriums, to recover this fixed point, and the algorithm exhibits a two-loop iteration structure. Exponential convergence is proved in this paper and we also verify our theoretical results by simple numerical examples.

Based on joint work with Fan Chen and Zhenjie Ren.

6.1 Introduction

Deep learning has achieved unprecedented success in numerous practical scenarios, including computer vision, natural language processing and even autonomous driving, which leverages deep reinforcement learning techniques [134, 96, 8]. Stochastic gradient algorithms (SGD) and their variants have been widely used to train neural networks, that is, to minimize networks' loss and thereby to fit the data available effectively [146, 131]. However, due to the complicated network structures and the non-convexity of typical optimization objectives, mathematical guarantees of convergence to the optimizer remain elusive. Recent studies on the insensibility of the number of neurons on one layer when it is sufficiently large [106], and the feasibility of interchanging the neurons on one layer [176, 203] both motivated the investigation of mean field regime. In practice, over-parameterized neural networks with a large number of neurons are commonly employed in order to achieve high performance [118]. This further motivates researchers to view neurons as random variables following a probability distribution and the summation over neurons as an expectation with respect to this distribution [212].

Another appealing approach to address the global convergence of such over-parameterized networks is through the *neural tangent kernel* (NTK) regime [125]. In this regime, it is believed that when the network width tends to infinity, the parameter updates, driven by stochastic gradient descent, do not significantly deviate from i.i.d Gaussian initialization, and these updates are called lazy training [219, 58]. As a result, training of neural networks can be depicted as regression with a fixed kernel given by linearization at initialization, leading to the exponential convergence [125]. By appropriate time rescaling, it is possible for the dynamics of the kernel method to track the SGD dynamics closely [162, 2]. Other studies, such as [73], explore the reproducing kernel Hilbert space and demonstrate that the gradient flow indeed converges to the kernel ridgeless regression with an adaptive kernel. Besides in [53], the researchers extend the definition of the kernel and show that the training with an appropriate regularizer also exhibits behaviors similar to the kernel method. However, the kernel behavior primarily manifests during the early stages of the training process, whereas the mean field model reveals and explains the longer-term characteristics [162]. Furthermore, another advantage of the mean field settings compared to NTK is the presence of feature learning, in contrast to the perspective of random feature [214, 95].

In the mean field limit where neurons become infinitely many, the dynamics of the neuron parameters under gradient descent can be understood as a gradient flow of measures in Wasserstein-2 space, providing a geometric interpretation of the learning algorithm. This flow is also described by a PDE system where the unknown is the density function of the measure. Well-posedness of the PDE system, discretization errors and finite-time propagation of chaos are studied in recent works [176, 162, 89, 7, 211]. On the other hand, extensive analysis has been conducted to investigate the convergence of such dynamics to their equilibrium. The convergence of gradient flows modeling shallow networks is studied in [57, 162, 117]; more recent works extend the gradient-flow formulation and study deep network structures [89, 176]. Sufficient conditions for the convergence under non-convex loss functions have been given in [176], and the discriminatory properties of the non-linear activation function have been exploited in [211, 203] to deduce the convergence.

In this paper, one key assumption is the convexity of the objective functional with respect to its measure-valued argument. This assumption has been exploited by many recent works. Notably, [178] have established the exponential convergence of the entropy-regularized problem in both discrete and continuous-time settings by utilizing the log-Sobolev inequality (LSI), following the observations in [179]. Additionally, [177] estimate the generalization error and prove a polynomial convergence rate by leveraging quadratic expansions of the loss function. [231] also prove polynomial convergence rates in different scenarios, where they add noise to the gradient descent and assume the activation and regularization functions are homogeneous.

With the existing convergence results on gradient flows for the mean field optimization problem in mind, the following question arises to us:

Do there exist dynamics other than gradient flows
that solve the (regularized) mean field optimization efficiently?

We believe the quest for its answer will not be wasted efforts, as it may lead to potentially highly performant algorithms for training neural networks, and also because the dynamics similar to that we consider in this paper have already found applications to various mean field problems.

We recall the classical fictitious play in game theory originally introduced by Brown [34] to learn Nash equilibriums. During the fictitious play, in each round of repeated games, each player optimally responds to the empirical frequency of actions taken by their opponents (hence the name). While the fictitious play does not necessarily converge in general cases [210], it does converge for zero-sum games [197] and potential games [164]. More recently, this method has been revisited in the context of mean field games [36, 104, 183, 144].

In this paper, we draw inspiration from the classical fictitious play and propose a similar algorithm, called *entropic fictitious play* (EFP), to solve mean field optimization problems emerging from the training of two-layer neural networks. Our algorithm shares a two-loop iteration structure with the particle dual average (PDA) algorithm, recently proposed by [179]. They estimated the computational complexity and conducted various numerical experiments for PDA to show its effectiveness in solving regularized mean field problems. However, PDA is essentially different from our EFP algorithm and their differences will be discussed in Sections 6.2 and 6.4.

6.2 Problem setting

Let us first recall how the (convex) mean field optimization problem emerges from the training of two-layer neural networks. While the universal representation theorem tells us that a two-layer network can arbitrarily well approximate the continuous function on the compact time interval [68, 13], it does not tell us how to find the optimal parameters. One is faced with the non-convex optimization problem

$$\min_{\beta_{n,i}\in\mathbb{R},\alpha_{n,i}\in\mathbb{R}^d,\gamma_{n,i}\in\mathbb{R}} \int_{\mathbb{R}\times\mathbb{R}^d} \ell\Big(y, \frac{1}{n}\sum_{i=1}^n \beta_{n,i}\varphi(\alpha_{n,i}\cdot z+\gamma_{n,i})\Big)\,\nu(dy\,dz), \tag{6.1}$$

where $\theta \mapsto \ell(y,\theta)$ is convex for every y, $\varphi : \mathbb{R} \to \mathbb{R}$ is a bounded, continuous and non-constant activation function, and ν is a measure of compact support representing the data. Denote the empirical law of the parameters m^n by $m^n = \frac{1}{n}\sum_{i=1}^n \delta_{(\beta_{n,i},\alpha_{n,i},\gamma_{n,i})}$. Then the neural network output can be written by

$$\frac{1}{n}\sum_{i=1}^n \beta_{n,i}\varphi(\alpha_{n,i}\cdot z+\gamma_{n,i}) = \int_{\mathbb{R}^{d+2}} \beta\varphi(\alpha\cdot z+\gamma)\, m^n(d\beta\, d\alpha\, d\gamma).$$

For technical reasons we may introduce a truncation function $h(\cdot)$ whose parameter is denoted by β as in [117]. To ease the notation we denote $x = (\beta,\alpha,\gamma) \in \mathbb{R}^{d+2}$ and $\hat{\varphi}(x,z) = h(\beta)\varphi(\alpha\cdot z+\gamma)$. Denote also by $\mathbb{E}^m = \mathbb{E}^{X\sim m}$ the expectation of the random variable X of law m. Now we relax the original problem (6.1) and study the mean field optimization problem over the probability measures,

$$\min_{m\in\mathcal{P}(\mathbb{R}^d)} F(m), \quad \text{where } F(m) := \int_{\mathbb{R}^d} \ell\big(y, \mathbb{E}^m[\hat{\varphi}(X,z)]\big)\,\nu(dy\,dz) \tag{6.2}$$

This reformulation is crucial, because the *potential* functional F defined above is convex in the space of probability measure. In this paper, as in [117, 163], we shall add a relative entropy term $H(m|g) := \int_{x\in\mathbb{R}^d} \log\frac{dm}{dg}(x)\, m(dx)$ in order to regularize the problem. The regularized problem then reads

$$\min_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma(m), \quad \text{where } V^\sigma(m) := F(m) + \frac{\sigma^2}{2}H(m|g). \tag{6.3}$$

Here we choose the probability measure g to be a Gibbs measure with energy function U, that is, the density of g satisfies $g(x) \propto \exp\big(-U(x)\big)$. It is worth noting that if a probability measure has finite entropy relative to the Gibbs measure g, then it is absolutely continuous with respect to the Lebesgue measure. Hence the density of m exists whenever $V^\sigma(m)$ is finite. In the following, we will abuse the notation and use the same letter to denote the density function of m.

Since F is convex, together with mild conditions, the first-order condition says that m^* is a minimizer of V^σ if and only if

$$\frac{\delta F}{\delta m}(m^*, x) + \frac{\sigma^2}{2}\log m^*(x) + \frac{\sigma^2}{2}U(x) = \text{constant}, \tag{6.4}$$

where $\frac{\delta F}{\delta m}$ is the linear derivative, whose definition is postponed to Assumption 6.1 below. Further, note that m^* satisfying (6.4) must be an invariant measure to the so-called mean field Langevin (MFL) diffusion:

$$dX_t = -\left(\nabla_x \frac{\delta F}{\delta m}(m_t, X_t) + \frac{\sigma^2}{2}\nabla_x U(X_t)\right) dt + \sigma\, dW_t, \quad \text{where } m_t := \mathrm{Law}(X_t).$$

In [117] it has been shown that the MFL marginal law m_t converges towards m^*, and this provides an algorithm to approximate the minimizer m^*.

The starting point of our new algorithm is to view the first-order condition (6.4) as a fixed pointed problem. Given $m \in \mathcal{P}(\mathbb{R}^d)$, let $\Phi(m)$ be the probability measure such that

$$\frac{\delta F}{\delta m}(m, x) + \frac{\sigma^2}{2}\log \Phi(m)(x) + \frac{\sigma^2}{2}U(x) = \text{constant}. \tag{6.5}$$

By definition, a probability measure m satisfies the first-order condition (6.4) if and only if m is a fixed point of Φ. Throughout the paper we shall assume that there exists at most one probability measure satisfying the first-order condition (equivalently, there exists at most one fixed point for Φ). This is true when the objective functional F is convex. Indeed, as the relative entropy $m \mapsto H(m|g)$ is strictly convex, the free energy $V^\sigma = F + \frac{\sigma^2}{2}H(\cdot|g)$ is also strictly convex and therefore admits at most one minimizer.

It remains to construct an algorithm to find the fixed point. Observe that $\Phi(m)$ defined in (6.5) satisfies formally

$$\Phi(m) = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathbb{E}^{X\sim\mu}\left[\frac{\delta F}{\delta m}(m, X)\right] + \frac{\sigma^2}{2}H(\mu|g), \tag{6.6}$$

that is, the mapping Φ is given by the solution to a variational problem, similar to the definition of Nash equilibrium. This suggests that we can adapt the classical fictitious play algorithm to approach the minimizer. In this context, $\Phi(m_t)$ is the "best response" to m_t in the sense of (6.6), and we define the evolution of the "empirical frequency" of the player's actions by

$$dm_t = \alpha\left(\Phi(m_t) - m_t\right) dt, \tag{6.7}$$

where α is a positive constant and should be understood as the learning rate. The Duhamel's formula for this equation reads

$$m_t = \int_0^t \alpha e^{-\alpha(t-s)}\, \Phi(m_s)\, ds + e^{-\alpha t}\, m_0,$$

so m_t is indeed a *weighted* empirical frequency of the previous actions m_0 and $\big(\Phi(m_s)\big)_{s\leqslant t}$.

We propose a numerical scheme corresponding to the entropic fictitious play described informally in Algorithm 3, which consists of inner and outer iterations. The inner iteration, described later in Algorithm 4 for a specific example, calculates an approximation of $\Phi(m_t)$ given the measure m_t. Note that we are sampling a classical Gibbs measure so various Monte Carlo methods can be used. The outer iterations let the measure evolve following the entropic fictitious play (6.7) with a chosen time step Δt.

Algorithm 3: Entropic fictitious play algorithm

Input: objective functional F, reference measure $g \propto \exp(-U)$, initial distribution m_0, time step Δt, interation times T.

```
1 for t = 0, Δt, 2Δt, ..., T − Δt do
      // Inner iteration
2     Sample Φ(m_{t+Δt}) ∝ exp(−(δF/δm)(m_t, x) − (σ²/2)U(x)) by Monte Carlo;
      // Outer iteration
3     Update m_{t+Δt} ← (1 − αΔt) m_t + αΔt Φ(m_n);
```

Output: distribution m_T.

6.2.1 Related works

Mean field optimization

In contrast to the entropy-regularized mean field optimization addressed by our EFP algorithm, the unregularized optimization has also been studied in recent works [57, 203, 211]. [89] developed a mean field framework that captures the feature evolution during multi-layer networks' training and analyze the global convergence for fully-connected neural networks and residual networks, introduced by [107]. Deep network settings have also been studied in [211, 175, 7, 184, 176].

Exponential convergence rate

The exponential convergence rate of the mean field Langevin dynamics has been shown in [178] by exploiting the log-Sobolev inequality, which critically relies on the non-vanishing entropic regularization. On the other hand, [56] has studied the annealed mean field Langevin dynamics, where the time steps decay following an $O\big((\log t)^{-1}\big)$ trend, and has shown the convergence towards the minimizer of the unregularized objective functional. In this paper, we will also prove an exponential convergence rate for our EFP algorithm and the precise statement can be found in Theorem 6.13. The convergence rate obtained solely depends on the learning rate, which can be chosen in a fairly arbitrary way. This seems to be an improvement over the LSI-dependent rate in [178, 56]. However, the arbitrariness is due to the fact that our theoretical result only addresses the outer iteration and assumes that the target measure of inner one can be perfectly sampled (see Algorithm 3), and our convergence rate can not be directly compared to the ones obtained by [178, 56]. However, the inner iteration aims to sample a Gibbs measure, which is a classical task for which various Monte Carlo algorithms are available. (see Remark 6.16).

Furthermore, we propose a "warm start" technique to alleviate the computational burden of the inner iterations (see Algorithm 4).

Particle dual averaging

Our entropic fictitious play algorithm shares similarities with the particle dual averaging algorithm introduced in [179]. PDA is an extension of regularized dual average studied in [174, 233], and can be considered the particle version of the dual averaging method designed to solve the regularized mean field optimization problem (6.3). The key feature shared by PDA and EFP is the two-loop iteration structure. In the PDA outer iteration, we calculate a moving average $\tilde{f}_n$ of the linear functional derivative of the objective $\frac{\delta F}{\delta m}$,

$$\tilde{f}_n = (1-\alpha\Delta t)\,\frac{\delta F}{\delta m}(\tilde{m}_{n-1},\cdot) + \alpha\Delta t\,\frac{\delta F}{\delta m}(\tilde{m}_{n-1},\cdot); \tag{6.8}$$

the measure $\tilde{m}_n$ is on the other hand updated by the inner iteration,

$$\tilde{m}_{n+1}(x) = \operatorname*{argmin}_{m\in\mathcal{P}(\mathbb{R}^d)} \mathbb{E}^m\big[\tilde{f}_n(x)\big] + \frac{\sigma^2}{2}H(m|g), \tag{6.9}$$

which can be calculated by a Gibbs sampler. While the PDA inner iteration (6.9) is identical to that of EFP, their outer iterations are distinctly different. The PDA outer iteration updates the linear derivatives $\frac{\delta F}{\delta m}(\tilde{m}_n,\cdot)$ by forming a convex combination, while the EFP outer iteration updates the measures by a convex combination, which serves as the first argument of the linear derivative $\frac{\delta F}{\delta m}(\cdot,\cdot)$. One disadvantage of PDA is that one needs to store the history of measures $(\tilde{m}_i)_{i=1}^n$ to evaluate $\tilde{f}_n$ in (6.8), which may lead to high memory usage in numerical simulations. Our EFP algorithm circumvents this numerical difficulty as the dynamics (6.7) corresponds to a birth-death particle system whose memory usage is bounded (see discussions in Section 6.4.2). As a side note, EFP and PDA coincide when the mapping $m \mapsto \frac{\delta F}{\delta m}(m,\cdot)$ is linear. This occurs when F is quadratic in m. For example, if F is defined by (6.2) with a quadratic loss, $\ell(y,\theta) = |y-\theta|^2$, then its functional derivative

$$\frac{\delta F}{\delta m}(m,x) = 2\int_{\mathbb{R}^d} \big(\mathbb{E}^m[\hat{\varphi}(X,z)] - y\big)\hat{\varphi}(x,z)\,\nu(dy\,dz)$$

is linear in m. Another difference is that the PDA outer iteration is updated with diminishing time steps (or equivalently, learning rates) $\Delta t = O\big(n^{-1}\big)$, which leads to the absence of exponential convergence, while EFP fixes the time step Δt and exhibits exponential convergence (modulo the errors from the inner iterations). Finally, the condition (A3) of [179] seems difficult to verify and our method does not rely on such an assumption.

6.2.2 Organization of paper

In Section 6.3 we state our results on the existence and convergence of entropic fictitious play. In Section 6.4 we provide a toy numerical experiment to showcase the feasibility of the algorithm for the training two-layer neural networks. Finally the proofs are given in Section 6.5 and they are organized in several subsections with a table of contents in the beginning to ease the reading.

6.3 Main results

Fix an integer $d > 0$ and a real number $p \geqslant 1$. Denote by $\mathcal{P}(\mathbb{R}^d)$ the set of the probability measures on $\mathbb{R}^d$ and by $\mathcal{P}_p(\mathbb{R}^d)$ the set of those with finite p-moment. We suppose the following assumption throughout the paper.

Assumption 6.1. 1. The mean field functional $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is non-negative and C^1, that is, there exists a continuous function, also called *functional linear derivative*, $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ such that for every m_0, $m_1 \in \mathcal{P}(\mathbb{R}^d)$,

$$F(m_1) - F(m_0) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_\lambda, x)(m_1 - m_0)\, m_\lambda(dx)\, d\lambda,$$

where $m_\lambda \coloneqq (1-\lambda)m_0 + \lambda m_1$. Moreover, there exists constants L_F, $M_F > 0$ such that for every m, $m' \in \mathcal{P}(\mathbb{R}^d)$ and for every x, $x' \in \mathbb{R}^d$,

$$\left| \frac{\delta F}{\delta m}(m, x) - \frac{\delta F}{\delta m}(m', x') \right| \leqslant L_F \big(\mathcal{W}_p(m, m') + |x - x'| \big), \tag{6.10}$$

$$\left| \frac{\delta F}{\delta m}(m, x) \right| \leqslant M_F. \tag{6.11}$$

2. The function $U : \mathbb{R}^d \to \mathbb{R}$ is measurable and satisfies

$$\int_{\mathbb{R}^d} \exp\big(-U(x)\big)\, dx = 1.$$

Moreover it satisfies

$$\operatorname*{ess\,inf}_{x \in \mathbb{R}^d} U(x) > -\infty \quad \text{and} \quad \liminf_{x \to \infty} \frac{U(x)}{|x|^p} > 0.$$

Given a function U satisfying Assumption 6.1, define the Gibbs measure g on $\mathbb{R}^d$ by its density $g(x) \coloneqq \exp(-U(x))$. In particular, given $m \in \mathcal{P}_p(\mathbb{R}^d)$, we can consider the relative entropy between m and g, [1]

$$H(m|g) = \int_{x \in \mathbb{R}^d} \log \frac{dm}{dg}(x)\, m(dx).$$

In this paper we consider the entropy-regularized optimization

$$\inf_{m \in \mathcal{P}(\mathbb{R}^d)} V^\sigma(m), \quad \text{where } V^\sigma(m) \coloneqq F(m) + \frac{\sigma^2}{2} H(m|g).$$

Our aim is to propose a dynamics of probability measures converging to the minimizer of the value function V^σ.

Proposition 6.2. *If Assumption 6.1 holds, then there exists at least one minimizer of V^σ, which is absolutely continuous with respect to the Lebesgue measure and belongs to $\mathcal{P}_p(\mathbb{R}^d)$.*

[1] The relative entropy is defined to be $+\infty$ whenever the integral is not well defined. Therefore, the relative entropy is defined for every measure in $\mathcal{P}(\mathbb{R}^d)$ and is always non-negative.

Given the result above, we can restrict ourselves to the space of probability measures of finite p-moments when we look for minimizers of the regularized problem V^σ. Before introducing the dynamics, let us recall the first-order condition for being a minimizer.

Proposition 6.3 (Proposition 2.5 of [117])**.** *Suppose Assumption 6.1 holds. If m^* minimizes V^σ in $\mathcal{P}(\mathbb{R}^d)$, then it satisfies the* first-order condition

$$\frac{\delta F}{\delta m}(m^*, \cdot) + \frac{\sigma^2}{2} \log m^*(\cdot) + \frac{\sigma^2}{2} U(\cdot) \textit{ is a constant Leb-a.e.,} \tag{6.12}$$

where $m^*(\cdot)$ denotes the density function of the measure m^*.

Conversely, if F is additionally convex, then every m^* satisfying (6.12) is a minimizer of V^σ and such a measure is unique.

Definition 6.4. For each $\mu \in \mathcal{P}(\mathbb{R}^d)$, define $G(\mu; \cdot) : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ by

$$G(\mu; m) = \mathbb{E}^{X \sim \mu}\left[\frac{\delta F}{\delta m}(m, X)\right]. \tag{6.13}$$

Furthermore, given $m \in \mathcal{P}(\mathbb{R}^d)$, we define a measure $\hat{m} \in \mathcal{P}(\mathbb{R}^d)$ by

$$\hat{m} = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathbb{R}^d)} G(\mu; m) + \frac{\sigma^2}{2} H(\mu | g), \tag{6.14}$$

whenever the minimizer exists and is unique.

Proposition 6.5. *Suppose Assumption 6.1 holds. The minimizer defined in (6.14) exists, is unique, and belongs to $\mathcal{P}_p(\mathbb{R}^d)$. This defines a mapping $\mathcal{P}_p(\mathbb{R}^d) \ni m \mapsto \hat{m} \in \mathcal{P}_p(\mathbb{R}^d)$, which we denote by Φ in the following.*

Since $\frac{\delta G}{\delta \mu}(\mu, x; m) = \frac{\delta F}{\delta m}(m, x)$, according to the first-order condition in Proposition 6.3, $\hat{m}$ must satisfy

$$\frac{\delta F}{\delta m}(m, \cdot) + \frac{\sigma^2}{2} \log \hat{m} + \frac{\sigma^2}{2} U \quad \text{is a constant Leb-a.e.} \tag{6.15}$$

Therefore, a probability measure m is a fixed point of the mapping Φ if and only if it satisfies the first-order condition (6.12). In particular, by Propositions 6.2 and 6.3, there exists at least one minimizer of V^σ, and it is a fixed point of the mapping Φ. On the other hand, if Φ admits only one fixed point, then it must be the unique minimizer of V^σ.

Given the definition of $\hat{m}$, the *entropic fictitious play* dynamics is the flow of measures $(m_t)_{t \geqslant 0}$ defined by

$$\frac{dm_t}{dt} = \alpha\,(\hat{m}_t - m_t). \tag{6.16}$$

This equation is understood in the sense of distributions a priori. We shall show that the entropic fictitious play converges towards the minimizer of V^σ under mild conditions.

Remark 6.6. Choosing the relative entropy to be the regularizer may seem arbitrary. It is motivated by the following two observations:

- If F is convex, the strict convexity of entropy ensures that the mapping Φ admits at most one fixed point.
- In numerical applications, one needs to sample the distribution $\hat{m}_t$ efficiently. Applying the entropic regularization, we can sample $\hat{m}_t$ by Monte Carlo methods since it is in the form of a Gibbs measure according to (6.14). See Section 6.4 for more details.

Definition 6.7 (Dynamical system per Definition 4.1.1 of [108])**.** Let $S[t]$ be a mapping from $\mathcal{W}_p$ to itself for every $t \geqslant 0$. We say the collection $(S[t])_{t\geqslant 0}$ is a *dynamical system* on $\mathcal{W}_p$ if

1. $S[0]$ is the identity on $\mathcal{W}_p$;
2. $S[t](S[t']m) = S[t+t']m$ for every $m \in \mathcal{P}_p(\mathbb{R}^d)$ and $t, t' \geqslant 0$;
3. for every $m \in \mathcal{P}_p(\mathbb{R}^d)$, $t \mapsto S[t]m$ is continuous;
4. for every $t \geqslant 0$, $m \mapsto S[t]m$ is continuous with respect to the topology of $\mathcal{W}_p$.

Proposition 6.8 (Existence and wellposedness of the dynamics)**.** *Suppose Assumption 6.1 holds. Let α be a positive real and let m_0 be in $\mathcal{P}_p(\mathbb{R}^d)$ for some $p \geqslant 1$. Then there exists a solution $(m_t)_{t\geqslant 0} \in C\big([0,+\infty); \mathcal{W}_p\big)$ to (6.16).*

When $p = 1$, the solution is unique and depends continuously on the initial condition. In other words, there exists a dynamical system $(S[t])_{t\geqslant 0}$ on $\mathcal{W}_1$ such that m_t defined by $m_t = S[t]m_0$ solves (6.16).

If additionally the initial value m_0 is absolutely continuous with respect to the Lebesgue measure, then the solution m_t admits density for every $t > 0$, and the densities $m_t(\cdot)$ solves (6.16) classically. That is to say, for every $x \in \mathbb{R}^d$ the mapping $t \mapsto m_t(x)$ is C^1 on $[0, +\infty)$ and the derivative satisfies

$$\frac{\partial m_t(x)}{\partial t} = \alpha\big(\hat{m}_t(x) - m_t(x)\big). \tag{6.17}$$

for every $t > 0$.

Now we study the convergence of the entropic fictitious play dynamics and to this end we introduce the following assumption.

Assumption 6.9.

1. The mapping $\Phi : \mathcal{P}_p(\mathbb{R}^d) \ni m \mapsto \hat{m} \in \mathcal{P}_p(\mathbb{R}^d)$ admits a unique fixed point m^*.
2. The initial value m_0 belongs to $\mathcal{P}_{p'}(\mathbb{R}^d)$ for some $p' > p$ and $H(m_0|g) < +\infty$.

Remark 6.10. Under Assumption 6.1, the first condition above is implied the convexity of F. Indeed, if F is convex, then the regularized objective V^σ reads $V^\sigma = F + H(\cdot|g)$ and is therefore strictly convex. So it admits a unique minimizer m^* in $\mathcal{P}_p(\mathbb{R}^d)$ and by our previous arguments $m^*$is also the unique fixed point of the mapping Φ.

Theorem 6.11 (Convergence in the general case)**.** *Let Assumptions 6.1 and 6.9 hold. If $(m_t)_{t\geqslant 0}$ is a flow of measures in $\mathcal{W}_p$ solving (6.16), then m_t converges to m^* in $\mathcal{W}_p$ when $t \to +\infty$, and for every $x \in \mathbb{R}^d$, $m_t(x) \to m^*(x)$ when $t \to +\infty$.*

Moreover, the mapping $t \mapsto V^\sigma(m_t)$ is differentiable with derivative

$$\frac{dV^\sigma(m_t)}{dt} = -\frac{\alpha\sigma^2}{2}\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big),$$

and it satisfies

$$\lim_{t\to+\infty} V^\sigma(m_t) = V^\sigma(m^*).$$

Given the convexity and higher differentiability of F, we also show that the convergence of $V^\sigma(m_t)$ is exponential.

Assumption 6.12. The mean-field function F is convex and C^2 with bounded derivatives. That is to say, there exists a continuous and bounded function $\frac{\delta^2 F}{\delta m^2}$: $\mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ such that it is the linear functional derivative of $\frac{\delta F}{\delta m}$.

Theorem 6.13. *Let Assumptions 6.1, 6.9 and 6.12 hold. Then we have for every $t\geqslant 0$,*

$$0 \leqslant V^\sigma(m_t) - \inf_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma(m) \leqslant \frac{\sigma^2}{2} H(m_0|\hat{m}_0)e^{-\alpha t}.$$

6.4 Numerical example

In this section we walk through the implementation of the entropic fictitious play in details by treating a toy example. Recall that in Algorithm 3 the measures are updated following the outer iteration

$$\frac{dm_t}{dt} = \alpha\,(\hat{m}_t - m_t),$$

and $\hat{m}_t = \Phi(m_t)$ is evaluated by the inner iteration.

6.4.1 Evaluation of Gibbs measure

Since $\hat{m}_t$ is a Gibbs measure corresponding to the potential $\frac{\delta F}{\delta m}(m_t,\cdot)+\frac{\sigma^2}{2}U$, it is the unique invariant measure of a Langevin dynamics under the following technical assumptions on F and U.

Assumption 6.14.
1. For all $m\in\mathcal{P}(\mathbb{R}^d)$, the function $\frac{\delta F}{\delta m}(m,\cdot):\mathbb{R}^d\to\mathbb{R}$ has a locally Lipschitz derivative, i.e. the intrinsic derivative of F, $DF(m,\cdot) \coloneqq \nabla\frac{\delta F}{\delta m}(m,\cdot)$ exists everywhere and is locally Lipschitz.
2. The function U is C^2, and there exists $\kappa>0$ such that $\big(\nabla U(x)-\nabla U(y)\big)\cdot(x-y)\geqslant\kappa(x-y)^2$ when $|x-y|$ is sufficiently large.

Proposition 6.15. *Suppose Assumptions 6.1 and 6.14 hold. Let m be a probability measure on $\mathbb{R}^d$. Then a probability measure $\hat{m}\in\mathcal{P}(\mathbb{R}^d)$ satisfies the condition (6.15) if and only if it is the unique stationary measure of the Langevin dynamics*

$$d\Theta_s = -\left(DF(m,\Theta_s)+\frac{\sigma^2}{2}\nabla U(\Theta_s)\right)ds+\sigma\,dW_s, \tag{6.18}$$

where W is a standard Brownian motion. Moreover, if $\mathrm{Law}(\Theta_0)\in\cup_{p>2}\mathcal{P}_p(\mathbb{R}^d)$, *then the marginal distributions* $\mathrm{Law}(\Theta_s)$ *converge in Wasserstein-2 distance towards the invariant measure.*

We refer readers to Theorem 2.11 of [117] for the proof of the proposition.

Remark 6.16. 1. There exist various Markov chain Monte Carlo (MCMC) methods for sampling Gibbs measures [5, 127]. Here in our inner iteration, we simulate the Langevin diffusion (6.18) by the simplest unadjusted Langevin algorithm (ULA) proposed in [181]. However, there are many other efficient MCMC methods for our aim. For example, we could employ the Metropolis-adjusted Langevin algorithms or the Hamiltonian Monte Carlo (HMC) methods based on an underdamped dynamics with fictitious momentum variables [172].

2. Exponential convergence in the sense of relative entropy for ULA proposed above is shown in [220], based on a log-Sobolev inequality condition for potential. There are also convergence results in the sense of the Wasserstein and total variation distance for Langevin Monte Carlo. For example, [81] prove Wasserstein convergence for ULA, [26, 54] prove respectively convergence in total variation and in Wasserstein distance for Hamiltonian Monte Carlo.

6.4.2 Simulation of entropic fictitious play

Now we explain our numerical scheme of the entropic fictitious play dynamics (6.16). First we approximate the probability distributions m_t by empirical measures of particles in the form

$$m_t = \frac{1}{N}\sum_{i=1}^{N} \delta_{\Theta_t^i},$$

where $\Theta_t^i \in \mathbb{R}^d$ encapsulates all the parameters of a single neuron in the network. In order to evaluate the Gibbs measure $\hat{m}_t$, we simulate a system of M Langevin particles using the Euler scheme for a long enough time S, i.e.,

$$\Theta_{t,s+\Delta s}^i = \Theta_{t,s}^i - \left(DF(m_t, \Theta_{t,s}^i) + \frac{\sigma^2}{2}\nabla U(\Theta_{t,s}^i) \right) \Delta s + \sigma\sqrt{\Delta s}\,\mathcal{N}_{t,s}^i, \tag{6.19}$$

for $1 \leqslant i \leqslant M$ and $s < S$, where $\mathcal{N}_{t,s}^i$ are independent standard Gaussian variables. We then set $\hat{m}_t$ equal to the empirical measure of the particles at the final time S, $(\Theta_{t,S}^i)_{1\leqslant i\leqslant M}$, i.e.,

$$\hat{m}_t \coloneqq \frac{1}{M}\sum_{i=1}^{M} \delta_{\Theta_{t,S}^i}.$$

To speed up the EFP inner iteration we adopt the following warm start technique. For each t, the initial value of the inner iteration $(\Theta_{t+\Delta t,0}^i)_{1\leqslant i\leqslant M}$ is chosen to be the final value of the previous inner iteration, i.e. $(\Theta_{t,S}^i)_{1\leqslant i\leqslant M}$. This approach exploits the continuity of the mapping Φ proved in Corollary 6.19: if Φ is continuous, the measures $\Phi(m_{t+\Delta t})$, $\Phi(m_t)$ should be close to each other as long as $m_{t+\Delta t}$, m_t are close, and this is expected to hold when the time step Δt is small. Hence this choice of initial value for the inner iterations should lead to less error in sampling the Gibbs measure $\hat{m}_t$.

Then we explain how to simulate the outer iteration. The naïve approach is to add particles to the empirical measures by

$$m_{t+\Delta t} = (1-\alpha\Delta t)\, m_t + \alpha\Delta t\, \hat{m}_t = \frac{1-\alpha\Delta t}{N}\sum_{i=1}^{N} \delta_{\Theta_t^i} + \frac{\alpha\Delta t}{N}\sum_{i=1}^{N} \delta_{\Theta_{t,S}^i}.$$

However, this leads to a linear explosion of the number of particles when $t \to +\infty$ as at each step it is incremented by M. To avoid this numerical difficulty, we view the EFP dynamics (6.16) as a birth-death process and kill $\lfloor \alpha \Delta t N \rfloor$ particles before adding the same number of particles that represents $\hat{m}_t$, calculated by the Gibbs sampler. In this way, the number of particles to keep remains bounded uniformly in time and the memory use never explodes.

6.4.3 Training a two-Layer neural network by entropic fictitious play

We consider the mean field formulation of two-layer neural networks in Section 6.1 with the following specifications. We choose the loss function ℓ to be quadratic: $\ell(y,\theta) = \frac{1}{2}|y-\theta|^2$, and the activation function to be the modified ReLU, $\varphi(t) = \max\big(\min(t,5),0\big)$. We also fix a truncation function h defined by

$$h(x) = \max\big(\min(x,5),-5\big).$$

In this case, the objective functional F reads

$$F(m) = \frac{1}{2K}\sum_{k=1}^{K}\Big(y_k - \mathbb{E}^m\big[h(\beta)\varphi(\alpha\cdot z_k+\gamma)\big]\Big)^2.$$

where (α,β,γ) is a random variable distributed as m and $(z_k,y_k)_{k=1}^K$ is the data set with z_k being the features and y_k being the labels. Finally we choose the reference measure g by fixing $U(x) = \frac{1}{2}x^2 + \text{constant}$, where the constant ensures that $\int g = \int \exp\big(-U(x)\big)\,dx = 1$. Under this choice, one can verify Assumptions 6.1, 6.9, 6.12, and the Langevin dynamics (6.19) for the inner iteration at time t reads

$$\begin{aligned}
d\beta_s &= \frac{1}{K}h'(\beta_s)\varphi(\alpha_s\cdot z_k+\gamma_s)\sum_{k=1}^{K}\Big(y_k - \mathbb{E}^{m_t}\big[h(\beta)\varphi(\alpha\cdot z_k+\gamma)\big]\Big)\,ds\\
&\qquad - \frac{\sigma^2}{2}\,\beta_s\,ds + \sigma\,dW_s^\beta\\
d\alpha_s &= \frac{1}{K}h(\beta_s)z_k\varphi'(\alpha_s\cdot z_k+\gamma_s)\sum_{k=1}^{K}\Big(y_k - \mathbb{E}^{m_t}\big[h(\beta)\varphi(\alpha\cdot z_k+\gamma)\big]\Big)\,ds\\
&\qquad - \frac{\sigma^2}{2}\,\alpha_s\,ds + \sigma\,dW_s^\alpha\\
d\gamma_s &= \frac{1}{K}h(\beta_s)\varphi'(\alpha_s\cdot z_k+\gamma_s)\sum_{k=1}^{K}\Big(y_k - \mathbb{E}^{m_t}\big[h(\beta)\varphi(\alpha\cdot z_k+\gamma)\big]\Big)\,ds\\
&\qquad - \frac{\sigma^2}{2}\,\gamma_s\,ds + \sigma\,dW_s^\gamma
\end{aligned}$$

where $W^{\{\alpha,\beta,\gamma\}}$ are independent standard Brownian motions in respective dimensions. The discretized version of this dynamics is then calculated on the interval $[0,S]$.

As a toy example, we approximate the 1-periodic sine function $z \mapsto \sin(2\pi z)$ defined on $[0,1]$ by a two-layer neural network. We pick $K = 101$ samples evenly distributed on the interval $[0,1]$, i.e. $z_k = \frac{k-1}{101}$, and set $y_k = \sin 2\pi z_k$ for $k = 1,\ldots,$ 101. The parameters for the outer iteration are

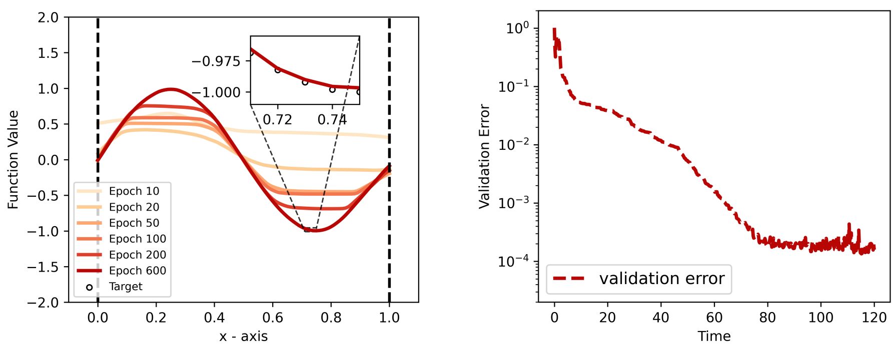
- time step $\Delta t = 0.2$,
- horizon $T = 120.0$,
- learning rate $\alpha = 1$,
- the number of neurons $N = 1000$,
- the initial distribution of neurons $m_0 = \mathcal{N}(0, 15^2)$.

For each t, we calculate the inner iteration (6.19) with the parameters:

- regularization $\sigma^2/2 = 0.0005$,
- time step $\Delta s = 0.1$,
- time horizon for the first step $S_{\text{first}} = 100.0$, and the remaining $S_{\text{other}} = 5.0$,
- the number of particles for simulating the Langevin dynamics $M = N = 1000$,

See Algorithm 4 for a detailed description.

We present our numerical results. We plot the learned approximative functions for different training epochs ($t/\Delta t = 10$, 20, 50, 100, 200, 600) and compare them to the objective in Figure 6.1(a). We find that in the last training epoch the sine function is well approximated. We also investigate the validation error, calculated from 1000 evenly distributed points in the interval $[0, 1]$, and plot its evolution in Figure 6.1(b). The final validation error is of the order of 10^{-4} and the whole training process consumes 63.02 seconds on the laptop (CPU model: i7-9750H). However, the validation error does not converge to 0, possibly due to the entropic regularizer added to the original problem.



(a) Approximated function value. (b) Validation error in training.

Figure 6.1: (a) The approximated function value at different time: the colors from shallow to deep represents the number of outer iterations processed, epoch 10, 20, 50, 100, 200, 600 respectively; (b) The validation error at different training epochs.

Algorithm 4: EFP with Langevin inner iterations

Input: objective function $F(\cdot)$, reference measure g with potential U, regularization parameter σ, initial distribution of parameter m_0, outer iterations time step Δt and horizon T, inner iterations time step Δs and horizon S, learning rate α, and number of particles in simulation N.

generate i.i.d. $\Theta_0^i \sim m_0, \;\; i = 1, \ldots, N$;
$(\Theta_{0,0}^i)_{i=1}^N \leftarrow (\Theta_0^i)_{i=1}^N$;
for $t = 0, \Delta t, 2\Delta t, \ldots, T - \Delta t$ **do**
    if $t = 0$ **then**
        $S \leftarrow S_{\text{first}}$;
    else
        $S \leftarrow S_{\text{other}}$;
    `// Inner iterations`
    for $s = 0, \Delta s, 2\Delta s, \ldots, S - \Delta s$ **do**
        generate standard normal variable $\mathcal{N}_{t,s}^i$;
        `// Update the inner particles by Langevin dynamics`
        for $i = 1, 2, \ldots, N$ **do**
            $\Theta_{t,s+\Delta s}^i \leftarrow \Theta_{t,s}^i - \left(DF(m_t, \Theta_t^i) + \frac{\sigma^2}{2}\nabla U(\Theta_{t,s}^i)\right)\Delta s + \sigma\sqrt{\Delta s}\,\mathcal{N}_{t,s}^i$;
    `// Outer iteration`
    $K \leftarrow \lfloor \alpha \Delta t N \rfloor$;
    choose uniformly K numbers from $\{1, \ldots, N\}$ and denote them by $(i_k)_{k=1}^K$;
    for $i = 1, 2, \ldots, N$ **do**
        if $i \in \{i_k\}_{k=1}^K$ **then**
            $\Theta_{t+\Delta t}^i \leftarrow \Theta_{t,S}^i$;
        else
            $\Theta_{t+\Delta t}^i \leftarrow \Theta_t^i$;
    `// Warm start for inner iterations`
    for $i = 1, 2, \ldots, N$ **do**
        $\Theta_{t+\Delta t,0}^i \leftarrow \Theta_{t,S}^i$;

Output: distribution $m_T = \frac{1}{N}\sum_{i=1}^N \delta_{\Theta_T^i}$.

6.5 Proofs

6.5.1 Proof of Propositions 6.2 and 6.5

Proof of Propositions 6.2 and 6.5. We only show Proposition 6.2 as the method is completely the same for the other proposition.

By Assumption 6.1 we have $\liminf_{x\to\infty} U(x)/|x|^p > 0$. Then we can find R, $c > 0$ such that $U(x) \geqslant c|x|^p$ for $|x| > R$. Choose a minimizing sequence $(m_n)_{n\in\mathbb{N}}$ in the sense that $V^\sigma(m_n) \searrow \inf_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma(m)$ when $n \to +\infty$. Then we have

$$\begin{aligned}
\sup_{n\in\mathbb{N}} V^\sigma(m_n) \geqslant H(m_n|e^{-U}) &= \int m_n(x)\big(\log m_n(x) + U(x)\big)\,dx \\
&= \bigg(\int_{|x|\leqslant R} + \int_{|x|>R}\bigg) m_n(x)\big(\log m_n(x) + U(x)\big)\,dx \\
&\geqslant -\frac{c_d R^d}{e} + \operatorname*{ess\,inf}_{|x|\leqslant R} U(x) + \int_{|x|>R} m_n(x)\big(\log m_n(x) + U(x)\big)\,dx \\
&\geqslant -\frac{c_d R^d}{e} + \operatorname*{ess\,inf}_{|x|\leqslant R} U(x) + \int_{|x|>R} m_n(x)\big(\log m_n(x) + c|x|^p\big)\,dx,
\end{aligned}$$

where the second inequality is due to $x\log x \geqslant -e^{-1}$ and c_d denotes the volume of the d-dimensional unit ball.

Define $\tilde{Z} = \int_{|x|>R} \exp\big(-c|x|^p/2\big)\,dx$ and denote by $\tilde{g}$ the probability measure

$$\tilde{g}(dx) = \frac{\mathbb{1}_{|x|>R}}{\tilde{Z}} \exp\Big(-\frac{c}{2}|x|^p\Big)\,dx$$

supported on $\{|x| > R\}$. Using the fact that the relative entropy is always nonnegative, we have

$$\begin{aligned}
&\int_{|x|>R} m_n(x)\big(\log m_n(x) + c|x|^p\big)\,dx \\
&\quad= \int_{|x|>R} m_n(x)\Big(\log m_n(x) + \frac{c}{2}|x|^p + \frac{c}{2}|x|^p\Big)\,dx \\
&\quad= H(m_n|\tilde{g}) - \log\tilde{Z}\int_{|x|>R} m_n(x)\,dx + \frac{c}{2}\int_{|x|>R} m_n(x)|x|^p\,dx \\
&\quad\geqslant -|\log\tilde{Z}| + \frac{c}{2}\int_{|x|>R} m_n(x)|x|^p\,dx.
\end{aligned}$$

Combining the two inequalities above, we obtain

$$\frac{c}{2}\int_{|x|>R} m_n(x)|x|^p\,dx \leqslant |\log\tilde{Z}| + \frac{c_d R^d}{e} - \operatorname*{ess\,inf}_{|x|<R} U(x) + \sup_{n\in\mathbb{N}} V^\sigma(m_n),$$

which implies

$$\sup_{n\in\mathbb{N}} \|m_n\|_p^p = \sup_{n\in\mathbb{N}} \int m_n(x)|x|^p\,dx < +\infty,$$

that is, the p-moment of the minimizing sequence is uniformly bounded. So the sequence $(m_n)_{n\in\mathbb{N}}$ is tight and $m_n \to m^*$ weakly for some $m^* \in \mathcal{P}(\mathbb{R}^d)$ along a subsequence. Applying the following lemma, whose proof is postponed, we obtain $m^* \in \mathcal{P}_p(\mathbb{R}^d)$.

Lemma 6.17 ("Fatou's lemma" for weak convergence of measure)**.** *Let X be a metric space, $f : X \to \mathbb{R}_+$ be nonnegative continuous function and $(m_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on X. If m_n converges to another probability measure m weakly, then*

$$\int_X f\,dm \leqslant \liminf_{n\to+\infty} \int_X f\,dm_n.$$

Since the relative entropy is weakly lower-semicontinuous, the entropy of m^* satisfies

$$H(m^*|g) \leqslant \liminf_{n\to+\infty} H(m_n|g).$$

We show the regular part satisfies $\lim_{n\to+\infty} F(m_n) = F(m^*)$. Indeed, by the definition of functional derivative, we have

$$\big|F(m_n) - F(m^*)\big| \leqslant \int_0^1 \bigg|\int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_{\lambda,n}, x)\,(m_n - m)(dx)\bigg|\,d\lambda$$

where $m_{\lambda,n} \coloneqq (1-\lambda)m_n + \lambda m$. For every $\lambda \in [0,1]$, we have

$$\begin{aligned}&\bigg|\int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_{\lambda,n}, x)\,(m_n - m^*)(dx)\bigg| \\ &\leqslant \bigg|\int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^*, x)\,(m_n - m^*)(dx)\bigg| + \int_{\mathbb{R}^d}\bigg|\frac{\delta F}{\delta m}(m_n, x) - \frac{\delta F}{\delta m}(m^*, x)\bigg|\,(m_n + m^*)(dx).\end{aligned}$$

Since $\frac{\delta F}{\delta m}(m^*,\cdot)$ is a bounded continuous function, the weak convergence $m_n \to m^*$ implies

$$\lim_{n\to+\infty}\bigg|\int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^*, x)\,(m_n - m^*)(dx)\bigg| = 0.$$

It remains to show the second term also converges to 0. Since the convergence $\frac{\delta F}{\delta m}(m_n, x) \to \frac{\delta F}{\delta m}(m^*, x)$ is uniform for $|x| < R$ for every $R > 0$, we have

$$\lim_{n\to+\infty} \int_{|x|\leqslant R}\bigg|\frac{\delta F}{\delta m}(m_n, x) - \frac{\delta F}{\delta m}(m^*, x)\bigg|\,(m_n + m^*)(dx) = 0.$$

Consequently,

$$\begin{aligned}\limsup_{n\to+\infty} &\int_{\mathbb{R}^d}\bigg|\frac{\delta F}{\delta m}(m_n, x) - \frac{\delta F}{\delta m}(m^*, x)\bigg|\,(m_n + m^*)(dx) \\ &= \limsup_{n\to+\infty}\bigg(\int_{|x|\leqslant R} + \int_{|x|>R}\bigg)\bigg|\frac{\delta F}{\delta m}(m_n, x) - \frac{\delta F}{\delta m}(m^*, x)\bigg|\,(m_n + m^*)(dx) \\ &\leqslant \limsup_{n\to+\infty}\int_{|x|>R}\bigg|\frac{\delta F}{\delta m}(m_n, x) - \frac{\delta F}{\delta m}(m^*, x)\bigg|\,(m_n + m^*)(dx) \\ &\leqslant M_F \limsup_{n\to+\infty}\int_{|x|>R}(m_n + m^*)(dx) \\ &= M_F \limsup_{n\to+\infty}\big(m^*(\{|x| > R\}) + m_n(\{|x| > R\})\big) = 0\end{aligned}$$

by tightness of the sequence $(m_n)_{n\in\mathbb{N}}$. Finally, using the boundedness

$$\bigg|\int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_{\lambda,n}, x)\,(m_n - m)(dx)\bigg| \leqslant 2M_F,$$

we can apply the dominated convergence theorem and show that when $n \to +\infty$,

$$\int_0^1 \left| \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_{\lambda,n}, x)\,(m_n - m)(dx) \right| d\lambda \to 0.$$

Summing up, we have obtained a measure $m^* \in \mathcal{P}_p(\mathbb{R}^d)$ such that

$$\begin{aligned} V^\sigma(m^*) &= F(m^*) + \frac{\sigma^2}{2} H(m^*) \\ &\leqslant \liminf_{n\to+\infty} F(m_n) + \frac{\sigma^2}{2} H(m_n) = \liminf_{n\to+\infty} V^\sigma(m_n) = \inf_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma(m). \end{aligned}$$

This completes the proof. □

Lemma 6.17. By the construction of Lebesgue integral, for every positive measure $\mu \in \mathcal{P}(X)$, we have

$$\int_X f\, d\mu = \sup_{M\geqslant 0} \int_X f \wedge M\, d\mu.$$

Therefore,

$$\begin{aligned} \int_X f\, dm &= \sup_{M\geqslant 0} \int_X f \wedge M\, dm \\ &= \sup_{M\geqslant 0} \liminf_{n\to+\infty} \int_X f \wedge M\, dm_n \\ &= \sup_{M\geqslant 0} \sup_{n} \inf_{k>n} \int_X f \wedge M\, dm_k \\ &\leqslant \sup_{n} \inf_{k>n} \sup_{M\geqslant 0} \int_X f \wedge M\, dm_k \\ &= \liminf_{n\to+\infty} \sup_{M\geqslant 0} \int_X f \wedge M\, dm_n \\ &= \liminf_{n\to+\infty} \int_X f\, dm_n, \end{aligned}$$

where the inequality is due to $\sup\inf \leqslant \inf\sup$. □

6.5.2 Proof of Proposition 6.8

We prove several technical results before proceeding to the proof of Proposition 6.8.

Proposition 6.18. *Suppose Assumption 6.1 holds. For every $m \in \mathcal{P}_p(\mathbb{R}^d)$, the measure $\hat{m}$ determined by*

$$\hat{m} = \frac{1}{Z_m} \exp\Big(-\frac{\delta F}{\delta m}(m, x) - U(x) \Big), \tag{6.20}$$

where Z_m is the normalization constant, is well defined and belongs to $\mathcal{P}_p(\mathbb{R}^d)$. Moreover, there exists constants c, C with $0 < c < 1 < C < +\infty$ such that for every $m \in \mathcal{P}_p(\mathbb{R}^d)$ and every $x \in \mathbb{R}^d$,

$$c e^{-U(x)} \leqslant \hat{m}(x) \leqslant C e^{-U(x)}, \tag{6.21}$$

Finally, there exists a constant $L > 0$ such that for every m, $m' \in \mathcal{P}_p(\mathbb{R}^d)$ and every $x \in \mathbb{R}^d$,

$$|\hat{m}(x) - \hat{m}'(x)| \leqslant L\mathcal{W}_p(m, m')e^{-U(x)}. \tag{6.22}$$

Proof. Using (6.11), we have

$$\exp\Big(-\frac{2}{\sigma^2}M_F - U(x)\Big) \leqslant \exp\Big(-\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m,x) - U(x)\Big) \leqslant \exp\Big(\frac{2}{\sigma^2}M_F - U(x)\Big), \tag{6.23}$$

and

$$\begin{aligned}\exp\Big(-\frac{2M_F}{\sigma^2}\Big)Z_0 &\\ = \int_{\mathbb{R}^d}\exp\Big(-\frac{2}{\sigma^2}M_F - U(x)\Big)dx \leqslant Z_m &\leqslant \int_{\mathbb{R}^d}\exp\Big(\frac{2}{\sigma^2}M_F - U(x)\Big)\,dx\\ &= \exp\Big(\frac{2M_F}{\sigma^2}\Big)Z_0,\end{aligned} \tag{6.24}$$

Thus $\hat{m}$ is well defined and (6.21) holds with constant $C = c^{-1} = \exp\big(4M_F\sigma^{-2}\big)$. Consequently,

$$\int_{\mathbb{R}^d}|x|^p\,\hat{m}(dx) \leqslant \int_{\mathbb{R}^d}|x|^p\hat{m}(x)\,dx \leqslant \int_{\mathbb{R}^d}|x|^pCe^{-U(x)}\,dx < +\infty,$$

that is, $\hat{m} \in \mathcal{P}_p(\mathbb{R}^d)$.

Meanwhile, using the elementary inequality $|e^x - e^y| \leqslant e^{x\vee y}|x-y|$, we have

$$\begin{aligned}&\left|\exp\Big(-\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m,x) - U(x)\Big) - \exp\Big(-\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m',x) - U(x)\Big)\right|\\ &\qquad\leqslant \frac{2}{\sigma^2}\exp\Big(\frac{2M_F}{\sigma^2}\Big)\mathcal{W}_p(m,m')\exp\big(-U(x)\big).\end{aligned}$$

Integrating the previous inequality with respect to x, we obtain

$$|Z - Z'| \leqslant \frac{2}{\sigma^2}\exp\Big(\frac{2M_F}{\sigma^2}\Big)\mathcal{W}_p(m,m')Z_0.$$

Using the bounds (6.23) and (6.24), we obtain the Lipschitz continuity (6.22). $\square$

The Lipschitz continuity (6.22) implies the Hölder continuity of $m \mapsto \hat{m}$.

Corollary 6.19. *Suppose Assumption 6.1 holds. Then the mapping $\Phi : \mathcal{P}_p(\mathbb{R}^d) \to \mathcal{P}_p(\mathbb{R}^d)$ is $1/p$-Hölder continuous.*

Before proving the corollary we show a lemma bounding the Wasserstein (coupling) distance between two probability measures by the L^∞ distance between their density functions.

Lemma 6.20. *Let (X, d) be a metric space and μ be a Borel probability measure on X. Consider the space of positive integrable functions with respect to μ,*

$$L^\infty_{+,1}(\mu) := \Big\{f : X \to \mathbb{R} \text{ Borel measurable} : f \geqslant 0, \int f\,d\mu = 1\Big\}.$$

Equip $L^\infty_{+,1}(\mu)$ with the usual L^∞ distance. Suppose for some $p \geqslant 1$ and some $x_0 \in X$, we have $C_{\mu,p} := \int_X d(x,x_0)^p \,\mu(dx) < +\infty$. Then there exists a constant $L_{\mu,p} > 0$ such that for every $f,\, g \in L^\infty_{+,1}(\mu)$,

$$\mathcal{W}_p(f\,\mu, g\,\mu) \leqslant L_{\mu,p} \|f-g\|^{1/p}_{L^\infty},$$

where $f\,\mu$ is the probability measure determined by $(f\,\mu)(A) := \int_A f\,d\mu$ and similarily for g.

Proof. Construct the following coupling π between $f\,\mu$, $g\,\mu$:

$$\begin{aligned}
\pi &:= \pi_1 + \pi_2,\\
\pi_1(dx\,dy) &:= (f\wedge g)(x)\,\mu_\Delta(dx\,dy),\\
\pi_2(dx\,dy) &:= \left(\int (f-g)_+(x)\mu(dx)\right)^{-1} (f-g)_+(x)(g-f)_+(y)\,\mu(dx)\,\mu(dy).
\end{aligned}$$

Here μ_Δ is the measure supported on the diagonal $\Delta := \{(x,x) : x\in X\} \subset X\times X$ such that $\mu_\Delta(A\times A) = \mu(A)$. One readily verifies that the projection mappings to the first and second variables, denoted by X, Y respectively, satisfy

$$\begin{aligned}
X_\# \pi_1 = Y_\# \pi_1 &= (f\wedge g)\,\mu,\\
X_\# \pi_2 &= (f-g)_+\,\mu,\\
Y_\# \pi_2 &= (g-f)_+\,\mu
\end{aligned}$$

Hence $X_\#\pi = f\,\mu$, $Y_\#\pi = g\,\mu$ and π is indeed a coupling between $f\,\mu$, $g\,\mu$.

By the definition of Wasserstein distance, we obtain

$$\begin{aligned}
\mathcal{W}_p(f\,\mu, g\,\mu)^p &\leqslant \int_{X\times X} d(x,y)^p \pi_1\,(dx\,dy) + \int_{X\times X} d(x,y)^p \pi_2\,(dx\,dy)\\
&= \left(\int (f-g)_+\,\mu\right)^{-1} \int_{X\times X} (f-g)_+(x)(g-f)_+(y) d(x,y)^p\,\mu(dx)\,\mu(dy).
\end{aligned}$$

Using triangle inequality $d(x,y)^p \leqslant C_p\big(d(x,x_0)^p + d(y,x_0)^p\big)$ and exchanging x, y, the last term is again bounded by

$$\begin{aligned}
&\frac{1}{\int (f-g)_+\,\mu} \int_{X\times X} C_p\big(d(x,x_0)^p + d(y,x_0)^p\big)(f-g)_+(x)(g-f)_+(y)\,\mu(dx)\,\mu(dy)\\
&\quad = \frac{C_p}{\int (f-g)_+\,\mu} \int_{X\times X} d(x,x_0)^p \Big[(f-g)_+(x)(g-f)_+(y)\\
&\qquad\qquad + (g-f)_+(x)(f-g)_+(y)\Big]\,\mu(dx)\,\mu(dy)\\
&\quad = C_p \int_X d(x,x_0)^p |f-g|(x)\,\mu(dx) \leqslant C_p C_{\mu,p} \|f-g\|_{L^\infty}.
\end{aligned}$$

The Hölder constant is then given by $L_{\mu,p} = (C_p C_{\mu,p})^{1/p}$. □

Remark 6.21. The Hölder exponent $1/p$ in the inequality is sharp. Consider the example: $\mu = \mathrm{Leb}_{[0,1]}$, $f = (1+\varepsilon)\mathbb{1}_{[0,1/2)} + (1-\varepsilon)\mathbb{1}_{[1/2,1]}$, $g = (1-\varepsilon)\mathbb{1}_{[0,\frac{1}{2})} + (1+\varepsilon)\mathbb{1}_{[1/2,1]}$. Then the $\mathcal{W}_p$ distance between $f\mu$, $g\mu$ is of order $\varepsilon^{1/p}$ when $\varepsilon \to 0$.

Proof of Corollary 6.19. Applying Lemma 6.20 with $\mu(dx) = e^{-U(x)}\,dx$, we obtain

$$\mathcal{W}_p(\hat m_1, \hat m_2) \leqslant L \left\| \frac{\hat m_1(x)}{e^{-U(x)}} - \frac{\hat m_2(x)}{e^{-U(x)}} \right\|_{L^\infty}^{1/p},$$

while by (6.22) we have

$$\left\| \frac{\hat m_1(x)}{e^{-U(x)}} - \frac{\hat m_2(x)}{e^{-U(x)}} \right\|_{L^\infty} \leqslant L\mathcal{W}_p(m_1, m_2).$$

The Hölder continuity follows. $\square$

Proof of Proposition 6.8. Step 1: Existence. We will use Schauder's fixed point theorem. To this end, fix $T > 0$, let $m_0 \in \mathcal{P}_p$ be the initial value and denote $X = C([0,T]; \mathcal{W}_p)$. Let $\boldsymbol{T} : X \to X$ be the mapping determined by

$$\boldsymbol{T}[m]_t \coloneqq \int_0^t \alpha e^{-\alpha(t-s)}\,\hat m_s\,ds + e^{-\alpha t}\,m_0 = \int_0^t \alpha e^{-\alpha(t-s)}\,\Phi(m_s)\,ds + e^{-\alpha t}\,m_0, \tag{6.25}$$

where $t \in [0,T]$. We verify indeed $\boldsymbol{T}[m] \in X$, i.e. $\boldsymbol{T}[m]_t \in \mathcal{P}_p$ for every $t \in [0,T]$, and $t \mapsto \boldsymbol{T}[m]_t$ is continuous with respect to $\mathcal{W}_p$. This first claim follows from the fact that $\boldsymbol{T}[m]_t$ is a convex combination of elements in $\mathcal{P}_p$, as we have shown $\hat m_s = \Phi(m_s) \in \mathcal{P}_p(\mathbb{R}^d)$. The second claim follows from

$$\begin{aligned}\mathcal{W}_p(\boldsymbol{T}[m]_{t+\delta}, \boldsymbol{T}[m]_t)^p &\leqslant \alpha \int_0^\delta e^{-\alpha(\delta-s)} \mathcal{W}_p(\hat m_s, m_t)^p\,ds \\ &\leqslant C(1 - e^{-\alpha\delta})\big(\sup\nolimits_{\hat m \in \operatorname{Im}\Phi} M_p(\hat m) + M_p(\boldsymbol{T}[m]_t)\big).\end{aligned} \tag{6.26}$$

Next we show the compactness of the mapping $\boldsymbol{T}$. Setting $t = 0$ in the previous equation and letting δ vary in $[0,T]$, we obtain

$$\sup_{m \in X} \sup_{t \in [0,T]} M_p\big(\boldsymbol{T}[m]_t\big) \leqslant C.$$

Plugging this back to (6.26), we have

$$\sup_{m \in X, 0 \leqslant t < t+\delta \leqslant T} \mathcal{W}_p\big(\boldsymbol{T}[m]_{t+\delta}, \boldsymbol{T}[m]_t\big) \leqslant C\delta^{1/p}. \tag{6.27}$$

From (6.11) one knows that $\operatorname{Im}\Phi$ forms a precompact set in $\mathcal{P}_p$, and since $X_t \coloneqq \{\boldsymbol{T}[m]_t : m \in X\}$ lies in the convex combination of $\operatorname{Im}\Phi$ and $\{m_0\}$, X_t is also precompact. Then by the Arzelà–Ascoli theorem, $\operatorname{Im}\boldsymbol{T} = \boldsymbol{T}[X]$ is a precompact set. In other words, $\boldsymbol{T}$ is a compact mapping. We use Schauder's theorem to conclude that $\boldsymbol{T}$ admits a fixed point, i.e. (6.16) admits at least one solution in X.

Step 2: Wellposedness when $p = 1$. The mapping Φ is Lipschitz in this case. The wellposedness follows from standard Picard–Lipschitz arguments.

Step 3: Pointwise solution. By definition, $\hat m_t$ admits the density function

$$\hat m_t(x) = \frac{1}{Z_t} \exp\Big(-\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t, x) - U(x)\Big),$$

where $Z_t \coloneqq \int_{\mathbb{R}^d} \exp\big(-\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x) - U(x)\big)\,dx$ is the normalization constant. The functional derivative $\frac{\delta F}{\delta m}(m_t, x)$ is continuous in t by the continuities of $t \mapsto m_t$ and

$m \mapsto \frac{\delta F}{\delta m}(m, x)$, and is bounded for every $t \geqslant 0$. By the dominated convergence theorem, both $\exp\left(-\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t, x) - U(x)\right)$ and Z_t are continuous in t and bounded. Hence $t \mapsto \hat{m}_t(x)$ is continuous and bounded uniformly in x. Suppose now the initial value m_0 has density $m_0(x)$. Define the density of m_t according to the Duhamel's formula (6.25):

$$m_t(x) \coloneqq \int_0^t \alpha e^{-\alpha(t-s)} \hat{m}_s(x)\, ds + e^{-\alpha t} m_0(x), \quad \text{for } x \in \mathbb{R}^d. \tag{6.28}$$

By definition $m_t(x)$ defined by (6.28) is indeed the density of m_t solving the time dynamics (6.16), and is automatically continuous in t. Since $\alpha e^{-\alpha(t-s)}\hat{m}_s(x)$ in (6.28) is continuous and bounded in s for every $t \geqslant 0$, the density $m_t(x)$ is C^1 in t and satisfies the pointwise equality (6.17). □

We also obtain a density bound that will be used in the following.

Corollary 6.22. *Suppose Assumption 6.1 holds. There exist constants c, $C > 0$, depending only on F and U, such that*

$$m_t(x) \geqslant (1 - e^{-\alpha t}) c e^{-U(x)}, \tag{6.29}$$

$$m_t(x) \leqslant (1 - e^{-\alpha t}) C e^{-U(x)} + e^{-\alpha t} m_0(x), \tag{6.30}$$

for every $x \in \mathbb{R}^d$.

Proof. For all $\hat{m} \in \operatorname{Im}\Phi$, we have

$$\hat{m}(x) \geqslant c e^{-U(x)}.$$

Then by the definition of density (6.28), we have

$$\begin{aligned} m_t(x) &\geqslant \int_0^t \alpha e^{-\alpha(t-s)} \hat{m}_s(x)\, ds \\ &\geqslant c e^{-U(x)} \int_0^t \alpha e^{-\alpha(t-s)} \hat{m}_s(x)\, ds \\ &= (1 - e^{-\alpha t}) c e^{-U(x)}. \end{aligned}$$

The proof for the upper bound is similar. □

6.5.3 Proof of Theorem 6.11

As it is important to our proof of Theorem 6.11, we single out the derivative in time result in the following proposition and prove it before tackling the other parts of the theorem.

Proposition 6.23. *Suppose Assumptions 6.1 and 6.9 holds, and let $(m_t)_{t\geqslant 0}$ be a solution to (6.16) in $\mathcal{W}_p$. Then for every $t > 0$,*

$$\frac{dV^\sigma(m_t)}{dt} = -\frac{\alpha\sigma^2}{2}\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big). \tag{6.31}$$

Before proving the proposition, we show a lemma on the uniform integrability of m_t and $\hat{m}_t$.

Lemma 6.24. *Fix $s > 0$. Under the conditions of the previous proposition, there exist integrable functions f, g such that for every $t \in [s, +\infty)$ and every $x \in \mathbb{R}^d$,*

$$g(x) \leqslant \log \frac{m_t(x)}{e^{-U(x)}} \big(\hat{m}_t(x) - m_t(x)\big) \leqslant f(x).$$

Proof. We first deal with the first term $\log \frac{m_t(x)}{e^{-U(x)}} \hat{m}_t(x)$. Using the bounds (6.29), (6.30) we have

$$\begin{aligned}\log \frac{m_t(x)}{e^{-U(x)}} \hat{m}_t(x) &\geqslant \log \frac{(1-e^{-\alpha t})ce^{-U(x)}}{e^{-U(x)}} \hat{m}_t(x) = \log\big((1-e^{-\alpha t})c\big)\hat{m}_t(x) \\ &\geqslant \log\big((1-e^{-\alpha s})c\big)\hat{m}_t(x) \geqslant \log\big((1-e^{-\alpha s})c\big) Ce^{-U(x)} =: g_1(x).\end{aligned}$$

Here we shrink the constant c if necessary so that $c < 1$ and in the last inequality the coefficient $\log\big((1-e^{-\alpha s})c\big)$ is negative. Now we upper bound $\log \frac{m_t(x)}{e^{-U(x)}} \hat{m}_t(x)$. We have

$$\begin{aligned}\log \frac{m_t(x)}{e^{-U(x)}} &\leqslant \log\biggl(e^{-\alpha t}\frac{m_0(x)}{e^{-U(x)}} + \int_0^t \alpha e^{-\alpha(t-s)} \frac{\hat{m}_s(x)}{e^{-U(x)}}\,ds\biggr) \\ &\leqslant \log\biggl(e^{-\alpha t}\frac{m_0(x)}{e^{-U(x)}} + C\int_0^t \alpha e^{-\alpha(t-s)}\,ds\biggr) \\ &= \log\biggl(e^{-\alpha t}\frac{m_0(x)}{e^{-U(x)}} + C(1-e^{-\alpha t})\biggr) \\ &\leqslant \log\big((1-e^{-\alpha t})C\big) + \frac{e^{-\alpha t}}{C(1-e^{-\alpha t})}\frac{m_0(x)}{e^{-U(x)}} \leqslant \log C + C_s \frac{m_0(x)}{e^{-U(x)}}.\end{aligned}$$

Here in the third inequality we used the elementary inequality $\log(x+y) \leqslant \log x + \frac{y}{x}$ for real x, y, and in the last line we maximize over $t \geqslant s$ and set $C_s = e^{-\alpha s}\big(C(1 - e^{-\alpha s})\big)^{-1}$. Therefore,

$$\begin{aligned}\log \frac{m_t(x)}{e^{-U(x)}} \hat{m}_t(x) &\leqslant \biggl(\log C + C_s\frac{m_0(x)}{e^{-U(x)}}\biggr)\hat{m}_t(x) \leqslant \biggl(\log C + C_s\frac{m_0(x)}{e^{-U(x)}}\biggr)Ce^{-U(x)} \\ &= \log C \cdot Ce^{-U(x)} + C_s C m_0(x) =: f_1(x).\end{aligned}$$

Now consider the second term $\log \frac{m_t(x)}{e^{-U(x)}} m_t(x)$. Applying Jensen's inequality to the Duhamel formula (6.28), we have

$$\begin{aligned}\log \frac{m_t(x)}{e^{-U(x)}} m_t(x) &\leqslant e^{-\alpha t}\log \frac{m_0(x)}{e^{-U(x)}} m_0(x) + \int_0^t \alpha e^{-\alpha(t-s)} \log \frac{\hat{m}_s(x)}{e^{-U(x)}}\hat{m}_s(x)\,dt \\ &\leqslant e^{-\alpha t}\log \frac{m_0(x)}{e^{-U(x)}} m_0(x) + \int_0^t \alpha e^{-\alpha(t-s)} \log C \cdot \hat{m}_s(x)\,dt \\ &\leqslant e^{-\alpha t}\log \frac{m_0(x)}{e^{-U(x)}} m_0(x) + \int_0^t \alpha e^{-\alpha(t-s)} \log C \cdot Ce^{-U(x)}\,dt \\ &\leqslant \biggl(\log \frac{m_0(x)}{e^{-U(x)}} m_0(x)\biggr)_+ + \log C \cdot Ce^{-U(x)} =: -g_2(x)\end{aligned}$$

In the second and third inequality we use consecutively the bound $\hat{m}(x) \leqslant Ce^{-U(x)}$ with $C > 1$. For the lower bound of the second term we note

$$\log \frac{m_t(x)}{e^{-U(x)}} m_t(x) = \log \frac{m_t(x)}{e^{-U(x)}} \cdot \frac{m_t(x)}{e^{-U(x)}} e^{-U(x)} \geqslant -\frac{1}{e}e^{-U(x)} =: -f_2(x)$$

The proof is complete by letting $f = f_1 + f_2$ and $g = g_1 + g_2$. □

Proof of Proposition 6.23. Thanks to the lemma above, we can apply the dominated convergence theorem to differentiate $t \mapsto V^\sigma(m_t)$ and obtain

$$\frac{dH(m_t)}{dt} = \alpha \int_{\mathbb{R}^d} \big(\log m_t(x) + U(x)\big)\big(\hat{m}_t(x) - m_t(x)\big)\,dx.$$

For the regular term $F(m_t)$, by the definition of functional derivative, we have

$$F(m_{t+\delta}) - F(m_t) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_{t+u\delta}, x)\big(m_{t+\delta}(x) - m_t(x)\big)\,dx\,du.$$

Applying again the dominated convergence theorem, the derivative reads

$$\begin{aligned}
\frac{dV^\sigma(m_t)}{dt} &= \alpha \int_{\mathbb{R}^d} \left(\frac{\delta F}{\delta m}(m_t, x) + \frac{\sigma^2}{2}\log m_t(x) + \frac{\sigma^2}{2}U(x)\right)\big(\hat{m}_t(x) - m_t(x)\big)\,dx \\
&= \alpha \int_{\mathbb{R}^d} \left(C_t + \frac{\sigma^2}{2}\log m_t(x) - \frac{\sigma^2}{2}\log \hat{m}_t(x)\right)\big(\hat{m}_t(x) - m_t(x)\big)\,dx \\
&= \alpha \int_{\mathbb{R}^d} \left(\frac{\sigma^2}{2}\log m_t(x) - \frac{\sigma^2}{2}\log \hat{m}_t(x)\right)\big(\hat{m}_t(x) - m_t(x)\big)\,dx \\
&= -\frac{\alpha\sigma^2}{2}\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big),
\end{aligned}$$

where in the second line we use the first-order condition for $\hat{m}_t$ and C_t is a constant that may depend on t. □

Remark 6.25. The result of Proposition 6.23 implies

- $\int_0^{+\infty} \big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big)\,dt < +\infty$;
- The derivative $\frac{dV^\sigma(m_t)}{dt}$ vanishes if and only if $m_t = \hat{m}_t$, i.e. the dynamics reaches a stationary point.

Proof of Theorem 6.11. Our strategy of proof is as follows. First we show that, by the (pre-)compactness of the flow $(m_t)_{t\geqslant 0}$ in a suitable Wasserstein space, the flow converges up to an extraction of subsequence. Then we prove by a monotonicity argument the convergence holds true without extraction. Finally we study the convergence of the density functions and prove the convergence of value function by the dominated convergence theorem.

According to the Duhamel's formula (6.25), the measure m_t is a (weighted) linear combination of the initial value m_0 and the best responses $\hat{m}_s$. Since there exists some $p' > p$ such that $m_0 \in \mathcal{P}_{p'}(\mathbb{R}^d)$, we obtain by the triangle inequality

$$\begin{aligned}
\|m_t\|_{p'}^{p'} &\leqslant e^{-\alpha t}\|m_0\|_{p'}^{p'} + (1 - e^{-\alpha t}) \sup_{0\leqslant s\leqslant t} \|\hat{m}_s\|_{p'}^{p'} \\
&\leqslant \|m_0\|_{p'}^{p'} + \sup_{m\in\mathcal{P}_p(\mathbb{R}^d)} \|\Phi(m)\|_{p'}^{p'} \leqslant \|m_0\|_{p'}^{p'} + C\int_{\mathbb{R}^d} x^{p'} e^{-U(x)}\,dx.
\end{aligned}$$

Thus the flow $(m_t)_{t\geqslant 0}$ in precompact in $\mathcal{P}_p(\mathbb{R}^d)$ and the set of limit points,

$$w(m_0) := \{m \in \mathcal{P}_p(\mathbb{R}^d) : \exists t_n \to +\infty \text{ such that } m_{t_n} \to m\},$$

is nonempty. We now show that $w(m_0)$ is the singleton $\{m^*\}$ and therefore $m_t \to m^*$ in $\mathcal{W}_p$. Pick $m \in w(m_0)$ and let $(t_n)_{n\in\mathbb{N}}$ be an increasing sequence such that

$t_n \to +\infty$ and $m_{t_n} \to m$. Extracting a subsequence if necessary, we may suppose $t_{n+1} - t_n \geqslant 1$ for $n \in \mathbb{N}$. Proposition 6.23 implies for every t, s such that $t > s \geqslant 0$,

$$V^\sigma(m_s) - V^\sigma(m_t) = \int_s^t \big(H(m_u|\hat{m}_u) + H(\hat{m}_u|m_u)\big)\,du.$$

Consequently,

$$\begin{aligned} V^\sigma(m_0) &\geqslant V^\sigma(m_{t_0}) - V^\sigma(m_{t_n}) \\ &\geqslant \sum_{k=0}^{n-1} \int_{t_k}^{t_k+1} \big(H(m_u|\hat{m}_u) + H(\hat{m}_u|m_u)\big)\,du \\ &\geqslant \sum_{k=0}^{n-1} \int_0^1 \big(H(m_{t_k+u}|\hat{m}_{t_k+u}) + H(\hat{m}_{t_k+u}|m_{t_k+u})\big)\,du. \end{aligned}$$

By taking $n \to +\infty$, we obtain

$$\sum_{k=0}^{n-1} \int_0^1 \big(H(m_{t_k+u}|\hat{m}_{t_k+u}) + H(\hat{m}_{t_k+u}|m_{t_k+u})\big)\,du < +\infty.$$

Therefore,

$$\begin{aligned} 0 &= \lim_{k\to+\infty} \int_0^1 \big(H(m_{t_k+u}|\hat{m}_{t_k+u}) + H(\hat{m}_{t_k+u}|m_{t_k+u})\big)\,du \\ &\geqslant \int_0^1 \liminf_{k\to+\infty} \big(H(m_{t_k+u}|\hat{m}_{t_k+u}) + H(\hat{m}_{t_k+u}|m_{t_k+u})\big)\,du \\ &= \int_0^1 \liminf_{k\to+\infty} \Big(H\big(S[u]m_{t_k}\big|\Phi(S[u]m_{t_k})\big) + H\big(\Phi(S[u]m_{t_k})\big|S[u]m_{t_k}\big)\Big)\,du \\ &= \int_0^1 \Big(H\big(S[u]m\big|\Phi(S[u]m)\big) + H\big(\Phi(S[u]m)\big|S[u]m\big)\Big)\,du. \end{aligned}$$

In the first inequality we applied Fatou's lemma, and in the last equality we used the convergence $m_{t_k} \to m$, the continuity of $S[u]$ and Φ, and the joint lower-semicontinuity of $(\mu,\nu) \mapsto H(\mu|\nu)$ with respect to the weak convergence of measures. Then we have

$$H\big(S[u]m\big|\Phi(S[u]m)\big) + H\big(\Phi(S[u]m)\big|S[u]m\big) = 0$$

for a.e. $u \in [0,1]$. Using again the lower-semicontinuity of relative entropy, we obtain

$$\begin{aligned} &H\big(m\big|\Phi(m)\big) + H\big(\Phi(m)\big|m\big) \\ &\qquad\qquad \leqslant \liminf_{u\to 0}\Big(H\big(S[u]m\big|\Phi(S[u]m)\big) + H\big(\Phi(S[u]m)\big|S[u]m\big)\Big) = 0. \end{aligned}$$

That is to say, as a probability measure $m = \Phi(m) = \hat{m}$. By our assumption Φ has unique fixed point m^*, therefore $m = m^*$ and $w(m_0)$ is equal to the singleton $\{m^*\}$.

Next we show that the convergence of the density function $m_t(\cdot) \to m^*(\cdot)$. Since $\frac{\sigma^2}{2}H(m^*) \leqslant V^\sigma(m^*) < +\infty$, the measure m^* has a density function, which we denote by $m^*(\cdot)$. The Duhamel's formula for density functions (6.28) yields

$$
\begin{aligned}
&\big|m_t(x) - m^*(x)\big| \\
&\quad \leqslant e^{-\alpha t}\big|m_0(x) - m^*(x)\big| + \int_0^t \alpha e^{-\alpha(t-s)}\big|\hat{m}_s(x) - m^*(x)\big|\,ds \\
&\quad \leqslant e^{-\alpha t}\big|m_0(x) - m^*(x)\big| + \int_0^t \alpha e^{-\alpha(t-s)} L\mathcal{W}_p(\hat{m}_s, m^*)e^{-U(x)}\,ds \\
&\quad = e^{-\alpha t}\big|m_0(x) - m^*(x)\big| + \int_0^t \alpha e^{-\alpha s} L\mathcal{W}_p(\hat{m}_{t-s}, m^*)e^{-U(x)}\,ds \\
&\quad = e^{-\alpha t}\big|m_0(x) - m^*(x)\big| + \int_0^{+\infty} \mathbb{1}_{s\leqslant t}\alpha e^{-\alpha s} L\mathcal{W}_p(\hat{m}_{t-s}, m^*)e^{-U(x)}\,ds.
\end{aligned}
$$

The integrand in the last integral is positive and upper-bounded by the integrable function

$$
\mathbb{1}_{s\leqslant t}\alpha e^{-\alpha s} L\mathcal{W}_p(\hat{m}_{t-s}, m^*)e^{-U(x)} \leqslant \alpha L \sup_{t\geqslant 0} \mathcal{W}_p(\hat{m}_t, m^*)e^{-\alpha s}e^{-U(x)},
$$

where $\sup_{t\geqslant 0}\mathcal{W}_p(\hat{m}_t, m^*) < +\infty$ because $(m_t)_{t\geqslant 0}$ is a continuous and convergent flow in $\mathcal{P}_p$. Hence by the dominated convergence theorem,

$$
\begin{aligned}
\lim_{t\to+\infty} &\int_0^{+\infty} \mathbb{1}_{s\leqslant t}\alpha e^{-\alpha s} L\mathcal{W}_p(\hat{m}_{t-s}, m^*)e^{-U(x)}\,ds \\
&= \int_0^{+\infty} \lim_{t\to+\infty} \mathbb{1}_{s\leqslant t}\alpha e^{-\alpha s} L\mathcal{W}_p(\hat{m}_{t-s}, m^*)e^{-U(x)}\,ds = 0,
\end{aligned}
$$

where $\lim_{s\to+\infty}\mathcal{W}_p(\hat{m}_s, m^*) = \lim_{s\to+\infty}\mathcal{W}_p\big(\Phi(m_s), m^*\big) = 0$ since $m_s \to m^*$ and Φ is continuous. As a result, $m_t(x) \to m^*(x)$ when $t \to +\infty$. We finally show the convergence of the value function. Note that, as in the proof of Proposition 6.23, the entropic term is doubly bounded by integrable functions

$$
-f_2(x) \leqslant m_t(x)\log\frac{m_t(x)}{e^{-U(x)}} \leqslant -g_2(x).
$$

Applying the dominated convergence theorem, we obtain

$$
\begin{aligned}
\lim_{t\to+\infty} H(m_t) = \lim_{t\to+\infty}\int_{\mathbb{R}^d} m_t(x)\log\frac{m_t(x)}{e^{-U(x)}}\,dx &= \int_{\mathbb{R}^d}\lim_{t\to+\infty} m_t(x)\log\frac{m_t(x)}{e^{-U(x)}}\,dx \\
&= \int_{\mathbb{R}^d} m^*(x)\log\frac{m^*(x)}{e^{-U(x)}}\,dx = H(m^*).
\end{aligned}
$$

The convergence in Wasserstein distance implies already $F(m_t) \to F(m^*)$. Therefore $\lim_{t\to+\infty} V^\sigma(m_t) = V^\sigma(m^*)$. □

6.5.4 Proof of Theorem 6.13

We again show some technical results before moving on to the proof of the theorem.

Lemma 6.26. *Suppose Assumptions 6.1 and 6.9 holds, and let m_t be a solution to (6.16). For every $t > 0$, we have*

$$0 \leqslant \int_{\mathbb{R}^d} \hat{m}_{t+\delta}(x)\left(\log \hat{m}_{t+\delta}(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t, x)\right) dx \\ - \int_{\mathbb{R}^d} \hat{m}_t(x)\left(\log \hat{m}_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t, x)\right) dx = O(\delta^{1/p})$$

when $\delta \to 0$.

Proof. Denote the quantity to bound by I. We write it as the sum of the following two terms:

$$\begin{aligned} I &= I_1 + I_2, \\ I_1 &= \int_{\mathbb{R}^d} \left(\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t, x) + \log \hat{m}_t(x) + U(x)\right)\big(\hat{m}_{t+\delta}(x) - \hat{m}_t(x)\big)\, dx = 0, \\ I_2 &= \int_{\mathbb{R}^d} \big(\log \hat{m}_{t+\delta}(x) - \log \hat{m}_t(x)\big)\hat{m}_{t+\delta}(x)\, dx. \end{aligned}$$

The term I_1 is zero because $\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t, x) + \log \hat{m}_t(x)$ is constant by the first-order condition. On the other hand, we have $I_2 = H(\hat{m}_{t+\delta}|\hat{m}_t) \geqslant 0$. Let us bound the other side. Since $\hat{m}_s(x) \geqslant ce^{-U(x)}$ holds for every $s \geqslant 0$, we have

$$\begin{aligned} &\big|\log \hat{m}_{t+\delta}(x) - \log \hat{m}_t(x)\big|\hat{m}_{t+\delta}(x) \\ &\quad\leqslant \frac{\hat{m}(x)}{\min\big\{\hat{m}_{t+\delta}(x), \hat{m}_t(x)\big\}}\big(m_{t+\delta}(x) - \hat{m}_t(x)\big) \\ &\quad\leqslant C\big(\hat{m}_{t+\delta}(x) - \hat{m}_t(x)\big) \\ &\quad\leqslant Ce^{-U(x)}\mathcal{W}_p(m_{t+\delta}, m_t) \\ &\quad\leqslant Ce^{-U(x)}\delta^{1/p}. \end{aligned}$$

Here we have used $\log \frac{x}{y} \leqslant \frac{|x-y|}{\min\{x,y\}}$ in the first inequality, (6.22) in the second inequality, and (6.27) in the last inequality. ☐

We need the following notion to treat the possibly non-differentiable relative entropy.

Definition 6.27. For a real function $f : (t-\varepsilon, t+\varepsilon) \to \mathbb{R}$ defined on a neighborhood of t, the set of its upper-differentials at t is

$$D^+ f(t) \coloneqq \left\{p \in \mathbb{R} : \limsup_{s \to t} \frac{f(s) - f(t) - p(s-t)}{|s-t|} \leqslant 0\right\}.$$

Lower-differentials are defined as $D^- f(t) \coloneqq -D^+(-f)(t)$.

Lemma 6.28. *Let $f : [a, b] \to \mathbb{R}$ be a function defined on a closed interval, continuous on its two ends a and b. If f has nonnegative lower-differentials on (a, b), i.e. for every $a < t < b$ there exists $p_t \in D^- f(t)$ with $p_t \geqslant 0$, then $f(b) \geqslant f(a)$.*

Proof. Since the interval $[a, b]$ is compact, for every $\varepsilon > 0$, we can find a finite sequence $a < x_1 < \ldots < x_n < b$ such that $f(x_{i+1}) - f(x_i) \geqslant -\varepsilon(x_{i+1} - x_i)$ with $x_1 < a+\varepsilon$ and $b < x_n+\varepsilon$. Thus we have $f(x_n)-f(x_1) \geqslant -\varepsilon(x_n-x_1)$. We conclude by taking the limit $\varepsilon \to 0$. ☐

Next we calculate the upper-differential of the relative entropy $t \mapsto H(m_t|\hat{m}_t)$.

Proposition 6.29. *Let Assumptions 6.1, 6.9 and 6.12 hold, and let $(m_t)_{t\geqslant 0}$ be a solution to (6.16) in $\mathcal{W}_p$. Then the relative entropy $H : t \mapsto H(m_t|\hat{m}_t)$ is continuous on $[0,+\infty)$, and for every $t>0$, the set of upper differentials $D^+H(t)$ is non-empty and there exists $p_t \in D^+H(t)$ such that*

$$p_t \leqslant -\alpha\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big).$$

Proof. Fix $t>0$. The relative entropy reads

$$\begin{aligned}
H_t \coloneqq H(m_t|\hat{m}_t) &= \int_{\mathbb{R}^d} m_t(x)\big(\log m_t(x) - \log \hat{m}_t(x)\big)\,dx \\
&= \int_{\mathbb{R}^d} m_t(x)\left(\log m_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right)dx \\
&\quad - \int_{\mathbb{R}^d} m_t(x)\left(\log \hat{m}_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right)dx \\
&= \int_{\mathbb{R}^d} m_t(x)\left(\log m_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right)dx \\
&\quad - \int_{\mathbb{R}^d} \hat{m}_t(x)\left(\log \hat{m}_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right)dx \\
&=: H_{1,t} - H_{2,t}.
\end{aligned}$$

In the second equality we can separate the integral into two parts because the integrand of the second term $m_t(x)\big(\log\hat{m}_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m,x)\big)$ is integrable as it is constant by the first-order condition. For the same reason, in the fourth equality we can replace m_t by $\hat{m}_t$ in the second term, as we are integrating against a constant and $m_t, \hat{m}_t$ have the same total mass 1.

Now we consider the difference $H_{t+\delta} - H_t = (H_{1,t+\delta} - H_{1,t}) - (H_{2,t+\delta} - H_{2,t})$. For the first part we have

$$\begin{aligned}
H_{1,t+\delta} - H_{1,t} &= H(m_{t+\delta}) - H(m_t) \\
&\quad + \frac{2}{\sigma^2}\int_{\mathbb{R}^d}\left(m_{t+\delta}(x)\frac{\delta F}{\delta m}(m_{t+\delta},x) - m_t(x)\frac{\delta F}{\delta m}(m_t,x)\right)dx. \\
&= \delta\int_{\mathbb{R}^d}\alpha\log\frac{m_t(x)}{e^{-U(x)}}(\hat{m}_t(x) - m_t(x))\,dx \\
&\quad + \frac{2\delta}{\sigma^2}\int_{\mathbb{R}^d}\alpha\big(\hat{m}_t(x) - m_t(x)\big)\frac{\delta F}{\delta m}(m_t,x)\,dx \\
&\quad + \frac{2\delta}{\sigma^2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} m_t(x)\frac{\delta^2 F}{\delta m^2}(m_t,x,y)\alpha\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy + o(\delta) \\
&= \alpha\delta\int_{\mathbb{R}^d}\left(\log m_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right)\big(\hat{m}_t(x) - m_t(x)\big)\,dx \\
&\quad + \frac{2\delta}{\sigma^2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} m_t(x)\frac{\delta^2 F}{\delta m^2}(m_t,x,y)\alpha\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy + o(\delta).
\end{aligned}$$

by Lemma 6.24 and dominated convergence theorem.

Next we calculate the second part:

$$\begin{aligned}H_{2,t+\delta} - H_{2,t} = {} & \frac{2}{\sigma^2}\int_{\mathbb{R}^d} \hat{m}_{t+\delta}(x)\left(\frac{\delta F}{\delta m}(m_{t+\delta},x) - \frac{\delta F}{\delta m}(m_t,x)\right) dx \\ & + \left[\int_{\mathbb{R}^d} \hat{m}_{t+\delta}(x)\left(\log \hat{m}_{t+\delta}(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right) dx \right. \\ & \quad \left. - \int_{\mathbb{R}^d} \hat{m}_t(x)\left(\log \hat{m}_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right) dx\right]\end{aligned}$$

For the first difference we use the expansion $m_{t+\delta} - m_t = \alpha\delta\,(\hat{m}_t - m_t) + o(\delta)$ and apply the dominated convergence theorem to obtain

$$\begin{aligned}&\frac{2}{\sigma^2}\int_{\mathbb{R}^d} \hat{m}_{t+\delta}(x)\left(\frac{\delta F}{\delta m}(m_{t+\delta},x) - \frac{\delta F}{\delta m}(m_t,x)\right) dx \\ &= \frac{2\alpha\delta}{\sigma^2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \hat{m}_{t+\delta}(x)\frac{\delta^2 F}{\delta m^2}(m_t,x,y)\alpha\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy + o(\delta) \\ &= \frac{2\alpha\delta}{\sigma^2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \hat{m}_t(x)\frac{\delta^2 F}{\delta m^2}(m_t,x,y)\alpha\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy + o(\delta)\end{aligned}$$

and the second difference is already treated in Lemma 6.26. Summing up, we have

$$H_{2,t+\delta} - H_{2,t} - \frac{2\alpha\delta}{\sigma^2}\iint \hat{m}_t(x)\frac{\delta^2 F}{\delta m^2}(m_t,x,y)\alpha\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy \geqslant o(\delta)$$

We have equally the bound on the other side: $H_{2,t+\delta} - H_{2,t} \leqslant O(\delta^{1/p})$.

Putting everything together, we have

$$\begin{aligned}&H_{t+\delta} - H_t \\ &\leqslant \alpha\delta\int\left(\log m_t(x) + U(x) + \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m_t,x)\right)\big(\hat{m}_t(x) - m_t(x)\big)\,dx \\ &\quad - \frac{2\alpha\delta}{\sigma^2}\iint \frac{\delta^2 F}{\delta m^2}(m_t,x,y)\big(\hat{m}_t(x) - m_t(x)\big)\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy + o(\delta) \\ &= \alpha\delta\int_{\mathbb{R}^d}\big(\log m_t(x) - \log \hat{m}_t(x)\big)\big(\hat{m}_t(x) - m_t(x)\big)\,dx \\ &\quad - \frac{2\alpha\delta}{\sigma^2}\iint \frac{\delta^2 F}{\delta m^2}(m_t,x,y)\big(\hat{m}_t(x) - m_t(x)\big)\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy + o(\delta) \\ &= -\alpha\delta\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big) \\ &\quad - \frac{2\alpha\delta}{\sigma^2}\iint \frac{\delta^2 F}{\delta m^2}(m_t,x,y)\big(\hat{m}_t(x) - m_t(x)\big)\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy + o(\delta).\end{aligned}$$

By the convexity of F. the double integral is positive, that is to say

$$\iint \frac{\delta^2 F}{\delta m^2}(m_t,x,y)\big(\hat{m}_t(x) - m_t(x)\big)\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy \geqslant 0.$$

For the other side we have $H_{t+\delta} - H_t \geqslant O(\delta^{1/p})$. Thus H_t is continuous and p_t defined by

$$\begin{aligned}p_t = {} & -\alpha\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big) \\ & - \frac{2\alpha}{\sigma^2}\iint \frac{\delta^2 F}{\delta m^2}(m_t,x,y)\big(\hat{m}_t(x) - m_t(x)\big)\big(\hat{m}_t(y) - m_t(y)\big)\,dx\,dy.\end{aligned}$$

is an upper-differential of $H(m_t|\hat{m}_t)$ and satisfies $p_t \leqslant -\alpha\big(H(m_t|\hat{m}_t)+H(\hat{m}_t|m_t)\big)$. □

Theorem of Theorem 6.13. By Proposition 6.23, we know

$$\frac{dV^\sigma(m_t)}{dt} = -\alpha\frac{\sigma^2}{2}\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big).$$

By Proposition 6.29, we find for every $t > 0$ an upper-differential $p_t \in D^+H(m_t|\hat{m}_t)$ such that

$$p_t \leqslant -\alpha\big(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\big).$$

Therefore,

$$\frac{dV^\sigma(m_t)}{dt} \geqslant \frac{\sigma^2}{2}p_t.$$

Since $\frac{dV^\sigma(m_t)}{dt} - p_t$ is a lower-differential of $V^\sigma(m_t) - H(m_t|\hat{m}_t)$, we apply Lemma 6.28 to the finite interval $[s,u]$ and obtain

$$V^\sigma(m_u) - V^\sigma(m_s) \geqslant \frac{\sigma^2}{2}\big(H(m_u|\hat{m}_u) - H(m_s|\hat{m}_s)\big). \tag{6.32}$$

It follows from Proposition 6.29 and Lemma 6.28 that $t \mapsto e^{\alpha t}H(m_t|\hat{m}_t)$ is non-increasing, and therefore,

$$H(m_t|\hat{m}_t) \leqslant H(m_0|\hat{m}_0)e^{-\alpha t}.$$

Taking the limit $u \to +\infty$ in (6.32), we obtain

$$\inf V^\sigma - V^\sigma(m_s) \geqslant 0 - \frac{\sigma^2}{2}H(m_s|\hat{m}_s) \geqslant -\frac{\sigma^2}{2}H(m_s|\hat{m}_s)e^{-\alpha t},$$

and the proof is complete. □

6.6 Conclusion

In this paper we proposed the entropic fictitious play algorithm that solves the mean field optimization problem regularized by relative entropy. The algorithm is composed of an inner and an outer iteration, sharing the same flavor with the particle dual average algorithm studied in [179], but possibly allows easier implementations. Under some general assumptions we rigorously prove the exponential convergence for the outer iteration and identify the convergence rate as the learning rate α. The inner iteration involves sampling a Gibbs measure and many Monte Carlo algorithms have been extensively studied for this task, so errors from the inner iterations are not considered in this paper. For further research directions, we may look into the discrete-time scheme to better understand the efficiency and the bias of the algorithm, and may also study the annealed entropic fictitious play (i.e., $\sigma \to 0$ when $t \to +\infty$) as well.

Chapter 7

Self-interacting approximation to McKean–Vlasov long-time limit: a Markov chain Monte Carlo method

7.1 Introduction

In this paper we develop a novel method to approximate the invariant measure of the non-degenerate McKean–Vlasov dynamics

$$\mathrm{d}X_t = b\big(\mathrm{Law}(X_t), X_t\big)\,\mathrm{d}t + \mathrm{d}B_t, \tag{7.1}$$

where B is a standard Brownian motion in $\mathbb{R}^d$. The McKean–Vlasov dynamics characterize the mean field limit of interacting particles, and they have widespread applications, encompassing fields such as granular materials [19, 23, 39], mathematical biology [129], statistical mechanics [161], and synchronization of oscillators [138]. More recently, there has been a growing interest in the role of such dynamics in the context of mean field optimization for training neural networks [163, 57, 117, 179, 178, 56, 63].

In order to simulate the invariant measure of (7.1), we turn to the corresponding N-particle approximation, i.e., the dynamics

$$\mathrm{d}X_t^i = b\bigg(\frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}, X_t^i\bigg)\,\mathrm{d}t + \mathrm{d}B_t^i, \qquad \text{for } i = 1, \ldots, N, \tag{7.2}$$

where $(B^i)_{1\leqslant i\leqslant N}$ are independent Brownian motions. It is expected that the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}$ of the N-particle system can approximate the McKean–Vlasov long-time limit when N and t are both large enough. For fixed N, to ensure control over the distance between the McKean–Vlasov dynamics in (7.1) and the N-particle system in (7.2), throughout the entire time horizon, the literature has proposed various uniform-in-time propagation of chaos results under different scenarios, see for example, [101, 80, 70, 142] and Chapter 1.

When the drift b does not depend on the marginal distribution, $\mathrm{Law}(X_t)$, the diffusion process is Markovian. Under general conditions, we can leverage Birkhoff's ergodic theorem [21] to approximate its invariant measure using the occupation measure

$$\overline{m}_t \coloneqq \frac{1}{t}\int_0^t \delta_{X_s}\,\mathrm{d}s,$$

as $t \to \infty$. In scenarios where the drift takes the form of gradient $b(x) = -\nabla U(x)$, this ergodic property of the Markov diffusion lays the groundwork for various Markov chain Monte Carlo methods, such as Metropolis adjusted Langevin algorithm [196, 195] and unadjusted Langevin algorithm [82]. Motivated by the Markovian ergodicity, the recent paper [75] studied the following self-interacting process:

$$\mathrm{d}X_t = b(\overline{m}_t, X_t)\,\mathrm{d}t + \mathrm{d}B_t, \tag{7.3}$$

where the dependence on the marginal distribution in McKean-Vlasov diffusion (7.1) is replaced by the occupation measure, that is, the empirical mean of the trajectory $(X_s)_{s\in[0,t]}$. In [75], the authors successfully demonstrated that, in the regime of weak interaction, where the dependence on the marginal distribution is sufficiently small, the occupation measures $(\overline{m}_t)_{t\geqslant 0}$ of the self-interacting process (7.3) also converge towards the invariant measure of (7.1) as $t \to \infty$. Remarkably, from a practical point of view, simulating the occupation measure of the self-interacting process (7.3) only requires a single particle, which distinguishes it from the conventional N-particle approximation (7.2). It is worth noting that the investigation into the long-time behavior of the self-interacting diffusions can be traced back to the pioneering works of Cranston and Le Jan [66] and Raimond [188].

Building upon the discovery in [75], this paper ventures into uncharted territory, where the mean field interaction need not to be inherently weak. We propose to study the self-interacting particle with exponentially decaying dependence on its trajectory:

$$\begin{aligned} \mathrm{d}X_t &= b(m_t, X_t)\,\mathrm{d}t + \mathrm{d}B_t, \\ \mathrm{d}m_t &= \lambda(\delta_{X_t} - m_t)\,\mathrm{d}t, \end{aligned} \tag{7.4}$$

where λ is a positive constant. Integrating the second equation of (7.4), we find

$$m_t = e^{-\lambda t}m_0 + \int_0^t \lambda e^{-\lambda(t-s)}\delta_{X_s}\,\mathrm{d}s,$$

that is to say, the measure m_t is an exponentially weighted occupation measure with emphasis on the recent past. The dynamics (7.4) is a time-homogeneous Markov process and we show its exponential ergodicity in the first part of the paper. Although the state space where the random variable (X_t, m_t) lives is infinite-dimensional, we have a non-degenerate noise in the X component and an almost sure contraction in the m component, which render such ergodicity possible. Under suitable conditions for the drift b, we show in Theorem 7.4 by a *reflection coupling* approach that the Markov process is contractive for a Wasserstein distance. This implies that the stationary measure ρ^λ for the Markov process exists, and is unique and globally attractive. Notably, the conditions that we impose on b do not imply the uniqueness of invariant measure for the McKean–Vlasov (7.1), and cover the case of the ferromagnetic Curie–Weiss model in Section 7.4. Here, we also remark that the exact weight for the measure m_t is not important and we work with a time-homogeneous Markov structure only for convenience. We could, for example, take the alternative weighting

$$\frac{\lambda}{1 - e^{-\lambda t}} \int_0^t e^{-\lambda(t-s)} \delta_{X_s} \, \mathrm{d}s$$

to remove the dependency on the initial value.

We then proceed to investigate properties of the stationary ρ^λ in Theorem 7.16. A crucial feature of the self-interacting process (7.4) is that when $\lambda \to 0$, the dynamics of the measure m_t becomes slow, while the rate of the X_t dynamics remains roughly unchanged. Suppose additionally that for some mean field energy functional $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, the drift is its negative intrinsic gradient:

$$b(m, x) = -D_m F(m, x) = -\nabla_x \frac{\delta F}{\delta m}(m, x).$$

The double time-scale structure allows us to speculate that $\mathrm{Law}(X_t)$ rapidly relaxes to the local-in-time equilibrium

$$\hat{m}_t := \frac{\exp\bigl(-2 \frac{\delta F}{\delta m}(m_t, x)\bigr) \, \mathrm{d}x}{\int \exp\bigl(-2 \frac{\delta F}{\delta m}(m_t, x')\bigr) \, \mathrm{d}x'}$$

so that the measure evolution is effectively described by

$$\partial_t m_t = \lambda(\hat{m}_t - m_t). \tag{7.5}$$

This dynamics is called *entropic fictitious play* in Chapter 6 and this point of view plays a key role in various literatures, notably the series of works of Benaïm, Ledoux and Raimond [15, 16, 17, 18] and the article of Kleptsyn and Kurtzmann [132]. The main novelty of our method is that we provide a *quantitative* justification of the intuition presented above, and we are no longer restricted to the case of two-body interaction potentials. More precisely, letting (X, m) be a random variable distributed as the stationary measure ρ^λ, we provide an entropy bound in Proposition 7.25 that measures in a way the distance between the conditional distribution $\mathrm{Law}(X|m)$ and $\hat{m}$, relying crucially on the log-Sobolev inequality for $\hat{m}$. This method requires unfortunately a finite-dimensional dependency of the mean field in the energy functional F, which we explain in Remark 7.17 in detail. The entropy bound, together with an inherent gradient structure of the dynamics, is then used in the rest of the proof of Theorem 7.16 to show that the random measure m solves approximately the stationary condition for the entropic fictitious play (7.5):

$$\hat{m} = m.$$

In the case that the energy F is convex, the equation above has a unique solution m_*, which is also the invariant measure for the McKean–Vlasov dynamics (7.1), and we show that the stationary measure ρ^λ is in fact close to $m_* \otimes \delta_{m_*}$ for small λ.

The self-interacting dynamics (7.4) can be also thought as an intermediate scheme between the entropic fictitious play (7.5), which corresponds to the limit $\lambda \to 0$, and the linear dynamics

$$\mathrm{d}X_t = b(\delta_{X_t}, X_t) \, \mathrm{d}t + \mathrm{d}B_t,$$

which corresponds to the limit $\lambda \to \infty$. From a computational point of view, the linear dynamics is easy to sample and relaxes rapidly, but in the long time does not yield the McKean–Vlasov's long-time limit. The entropic fictitious play reaches high precision in the long time, but at each time step, it requires usually a costly Monte Carlo run to sample the Gibbs measure $\hat{m}_t$. The self-interacting dynamics lies exactly in between by mixing the two time scales.

As a final note on the terminology, although the words "stationary" and "invariant" have almost identical meanings in the context of stochastic process, we always say "invariant measure" when referring to the McKean–Vlasov process (7.1), and "stationary measure" when referring to the self-interacting process (7.4). We hope this artificial distinction would reduce possible confusions for the readers.

The rest of the paper is organized as follows. The main results are stated in Section 7.2. Before moving to their proofs, we apply our results to the training of a two-layer neural network in Section 7.3 and to a Curie–Weiss model for ferromagnets in Section 7.4. Ergodicity of the self-interacting dynamics, i.e., Theorem 7.4, is proved in Section 7.5. In Section 7.6, we prove Theorem 7.16, which characterizes the stationary measure of the self-interacting process. Finally, a technical result and its proof, and the numerical algorithm are included in the appendices.

7.2 Main results

We state and discuss our main results in this section. First, we study the contractivity of the self-interacting process (7.4), and in particular, the exponential convergence to its unique stationary measure is obtained. Then, focusing on the gradient case, we quantify the distance between the self-interacting stationary measure and the corresponding McKean–Vlasov invariant measure. Finally, applying both the results, we propose an annealing scheme so that the self-interacting dynamics converges to the McKean–Vlasov invariant measure.

To avoid extra assumptions on the drift functional b, we will always assume the existence and the uniqueness of strong solution to the self-interacting process (7.4) without explicitly mentioning it in the rest of the paper.

Assumption. Given any filtered probability space supporting a Brownian motion $(B_t)_{t\geqslant 0}$ and satisfying the usual conditions, for any initial conditions (X_0, m_0) measurable to the initial σ-algebra and taking value in $\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$, there exists a unique adapted stochastic process $(X_t, m_t)_{t\geqslant 0}$ such that for all $t \geqslant 0$,

$$
\begin{aligned}
X_t &= \int_0^t b(m_s, X_s)\,\mathrm{d}s + B_t + X_0,\\
m_t &= \lambda \int_0^t (\delta_{X_s} - m_s)\,\mathrm{d}s + m_0.
\end{aligned}
$$

One may easily find various sufficient conditions for the assumption above. For example, if $b : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ is W_1-Lipschitz continuous in measure and Lipschitz continuous in space, then by Cauchy–Lipschitz arguments, we know that the assumption is satisfied.

7.2.1 Contractivity of the self-interacting diffusion

We first present the results on the contractivity of the self-interacting dynamics (7.4), from which follows the convergence to its unique stationary measure. We start with two basic definitions. First, we define moduli of continuity.

Definition 7.1 (Modulus of continuity)**.** We say that $\omega : [0,\infty) \to [0,\infty)$ is a modulus of continuity if it satisfies the following properties:

- $\omega(0) = 0$;
- ω is continuous and non-decreasing;
- for all h, $h' \geqslant 0$, we have $\omega(h+h') \leqslant \omega(h) + \omega(h')$.

Note that a modulus of continuity necessarily has at most linear growth according to our definition. We also introduce the notion of semi-monotonicity following Eberle [83].

Definition 7.2 (Semi-monotonicity)**.** We say that $\kappa : (0,\infty) \to \mathbb{R}$ is a semi-monotonicity function for a vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ if

$$\big(v(x) - v(x')\big) \cdot (x - x') \leqslant -\kappa(|x-x'|)|x-x'|^2$$

holds for every x, $x' \in \mathbb{R}^d$ with $x' \neq x$. We say κ is a *uniform* semi-monotonicity function of a family of vector fields if it is a semi-monotonicity function of each member.

In this subsection, we impose the following assumption on the drift of the McKean–Vlasov dynamics (7.1).

Assumption 7.3. The drift b satisfies the following conditions:

1. For any fixed $m \in \mathcal{P}(\mathbb{R}^d)$, the vector field $x \mapsto b(m,x)$ is uniformly equicontinuous and has a uniform semi-monotonicity function κ_b, given by $\kappa_b(x) = \kappa_0 - M_b/x$ for some $\kappa_0 > 0$ and $M_b \geqslant 0$.

2. There exist a bounded modulus of continuity $\omega : [0,\infty) \to [0, M_\omega]$ and a constant $L \geqslant 0$ such that

$$|b(m,x) - b(m',x)| \leqslant L W_\omega(m,m')$$

for every m, $m' \in \mathcal{P}(\mathbb{R}^d)$ and every $x \in \mathbb{R}^d$. Here W_ω is the Wasserstein distance

$$W_\omega(m,m') = \inf_{\pi \in \Pi(m,m')} \int \omega(|x-x'|)\pi(\mathrm{d}x\,\mathrm{d}x').$$

Using reflection coupling, we derive the following result.

Theorem 7.4. *Suppose Assumption 7.3 hold. Let $(X_t, m_t)_{t\geqslant 0}$, $(X'_t, m'_t)_{t\geqslant 0}$ be two processes following the dynamics (7.4) for some $\lambda > 0$ such that the first marginals of their initial values X_0, X'_0 have finite first moments. Define the following metric on $\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$:*

$$d_\lambda\big((x,m),(x',m')\big) = |x-x'| + \frac{2L}{\lambda} W_\omega(m,m')$$

and denote the corresponding Wasserstein distance on $\mathcal{P}_1\big(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\big)$ *by* W_{d_λ}*. Then, we have*

$$W_{d_\lambda}\big(\mathrm{Law}(X_t, m_t), \mathrm{Law}(X'_t, m'_t)\big) \leqslant Ce^{-ct} W_{d_\lambda}\big(\mathrm{Law}(X_0, m_0), \mathrm{Law}(X'_0, m'_0)\big),$$

where the constants C, c *are given by*

$$C = 1 + \frac{2M}{\sqrt{K_0}} \exp\biggl(\frac{M^2}{4K_0}\biggr),$$
$$c = \biggl(\frac{1}{K_0} + \frac{2M}{K_0^{3/2}} \exp\biggl(\frac{M^2}{4K_0}\biggr)\biggr)^{-1}$$

for $M = M_b + 2LM_\omega$ *and* $K_0 = \min\bigl(\kappa_0, \frac{\lambda}{2}\bigr)$*.*

The proof of Theorem 7.4 is postponed to Section 7.5.

Remark 7.5 (On the assumption). The first condition on the semi-monotonicity of the vector field $x \mapsto b(m, x)$ is stronger than those used in [83], in that we require a more gentle singularity in $\kappa(x)$ for x close to 0. This is because in this work, we are concerned with a good convergence rate when the parameter $\lambda \to 0$ (see the following remark for more discussions) and it will become clear in the proof that this stronger requirement on the semi-monotonicity is necessary for our purpose. Nevertheless, this condition is not too difficult to fulfill. Upon defining $b_0(x) = -\kappa_0 x$ and $b_1(m, x) = b(m, x) - b_0(x)$, the first condition of Assumption 7.3 is equivalent to

$$\bigl(b_1(m, x) - b_1(m, x'), x - x'\bigr) \leqslant M_b,$$

and this holds true when $\sup_{(m,x)\in\mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d} |b_1(m, x)| \leqslant M_b/2$ in particular.

Remark 7.6 (Rate of convergence). We discuss the rate of convergence across three parameter ranges.

In the first scenario, when the drift b is κ_0-strongly monotone for some $\kappa_0 > 0$, i.e., $M_b = 0$, and when there is no mean field interaction ($L = 0$), we have $M = 0$ and $K_0 = \min\bigl(\kappa_0, \frac{\lambda}{2}\bigr)$. Consequently, $C = 1$ and $c = \min\bigl(\kappa_0, \frac{\lambda}{2}\bigr)$. It is worth noting that under these conditions, the component X exhibits exponential contraction with a rate of κ_0, while m contracts at a rate no greater than λ. In this case, the best contraction rate for the joint process is $\min(\kappa_0, \lambda)$. Thus our method yields a contraction rate that remains at least half of the optimal one.

In the second scenario, when λ is small but non-zero (with self-interaction), we obtain $c \sim 2M\lambda^{3/2} \exp(-M^2/2\lambda)$ and $C \sim 2M\lambda^{-1/2} \exp(-M^2/2\lambda)$. We note that such exponentially slow convergence also arises in the kinetic Langevin process as the damping parameter approaches zero; see Eberle, Guillin and Zimmer [84, Section 2.6] for further discussion.

Finally, for $\lambda > 2\kappa_0$, the contractivity constants C, c become independent of λ, consistent with the intuition that the self-interacting diffusion becomes linear in the large λ limit.

Now we discuss a few examples satisfying Assumption 7.3.

Example 7.7 (Two-body interaction). Consider $b(m, x) = b_0(x) + b_1(m, x)$, where

$$b_1(m, x) = \int K(x, x') m(\mathrm{d}x').$$

Suppose furthermore that

$$\sup_{x\in\mathbb{R}^d} \|K(x,\cdot)\|_{W^{1,\infty}} = \sup_{x\in\mathbb{R}^d} \max\bigl(\|K(x,\cdot)\|_{L^\infty}, \|\nabla K(x,\cdot)\|_{L^\infty}\bigr) \leqslant M,$$

that is to say, the mapping $y \mapsto K(x,y)$ is M-bounded and M-Lipschitz continuous for every x. Thus, we have

$$|b(m,x) - b(m',x)| \leqslant \left|\int K(x,x')(m-m')(\mathrm{d}x')\right| \leqslant MW_\omega(m,m')$$

for the modulus of continuity $\omega(x) = \min(x,2)$. Therefore Assumption 7.3 is satisfied once b_0 is uniformly Lipschitz and has a semi-monotonicity function $\kappa_0(x) = \kappa_0 - M_1/x$ for some $\kappa_0 > 0$ and $M_1 \geqslant 0$.

We can generalize the example above to drifts that depend on the measure in a non-linear way.

Example 7.8 ($\mathcal{C}^1$ functional). Suppose $m \mapsto b(m,x)$ is $\mathcal{C}^1$ differentiable in the sense that there exists a continuous and bounded mapping $\frac{\delta b}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$b(m,x) - b(m',x) = \int_0^1 \int \frac{\delta b}{\delta m}\bigl((1-t)m + tm', x, x'\bigr)(m-m')(\mathrm{d}x')\,\mathrm{d}t$$

for all m, $m' \in \mathcal{P}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$. If $\sup_{m,x}\bigl\|\frac{\delta b}{\delta m}(m,x,\cdot)\bigr\|_{W^{1,\infty}}$ is finite and the vector fields $x \mapsto b(m,x)$ are uniformly Lipschitz and share a uniform semi-monotonicity function $\kappa_0(x) = \kappa_0 - M_1/x$ for some $\kappa_0 > 0$ and $M_1 \geqslant 0$, then by the same argument as in Example 7.7, Assumption 7.3 is satisfied.

We now examine the stationary measure of the self-interacting process (7.4).

Definition 7.9. We call a probability measure $P \in \mathcal{P}\bigl(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\bigr)$ stationary to the self-interacting diffusion (7.4) if the stochastic process $(X_t, m_t)_{t\geqslant 0}$ with initial value $\mathrm{Law}(X_0, m_0) = P$ satisfies $\mathrm{Law}(X_t, m_t) = P$ for all $t \geqslant 0$.

The definition above makes sense since we have assumed the existence and uniqueness of strong solution.

By the Banach fixed point theorem in the metric space $\mathcal{P}_1\bigl(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\bigr)$ and standard arguments, Theorem 7.4 implies the existence and uniqueness of stationary measure of the self-interacting process (7.4).

Corollary 7.10. *Under Assumption 7.3, for every $\lambda > 0$, there exists a unique stationary measure of the self-interacting diffusion (7.4) in $\mathcal{P}\bigl(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\bigr)$ whose first marginal distribution on $\mathbb{R}^d$ has finite first moment.*

Finally, we note that although Theorem 7.4, along with Corollary 7.10, implies that the self-interacting process (7.4) converges to its stationary measure exponentially, its assumptions are not sufficient to establish the uniqueness of invariant measure of the McKean–Vlasov process (7.1), as illustrated by the Curie–Weiss example in Section 7.4. So in general, there is no hope that the self-interacting stationary measure approximates the McKean–Vlasov invariant measure.

7.2.2 Stationary measure in the gradient case

In this subsection, we study the stationary measure of the self-interacting dynamics (7.4), provided that the drift b is the negative intrinsic gradient of a finite-dimensional mean field functional, whose precise meaning will be explained in the following. In particular, we aim at proving that, in this case, the stationary measure of (7.4) provides an approximation to the invariant measure of the McKean–Vlasov dynamics (7.1). We fix a positive λ in this subsection.

The first assumption in the subsection is that the drift b corresponds to a gradient descent whose dependency in the mean field is finite-dimensional.

Assumption 7.11 (Finite-dimensional mean field)**.** For a closed convex set $\mathcal{K} \subset \mathbb{R}^D$, there exists a function

$$\ell = (\ell^1, \ldots, \ell^D) \in \mathcal{C}^1(\mathbb{R}^d; \mathcal{K})$$

whose gradient is of at most linear growth, and a function $\Phi \in \mathcal{C}^2(\mathbb{R}^D; \mathbb{R})$ whose Hessian $\nabla^2\Phi$ is bounded, such that the drift term b reads

$$\begin{aligned} b(m,x) &= -\nabla\Phi(\langle \ell, m\rangle) \cdot \nabla\ell(x) = -\nabla\Phi\biggl(\int \ell(x)m(\mathrm{d}x)\biggr) \cdot \nabla\ell(x) \\ &= -\sum_{\nu=1}^{D} \nabla_\nu\Phi\biggl(\int \ell(x)m(\mathrm{d}x)\biggr)\nabla\ell^\nu(x). \end{aligned}$$

In other words, for the mean field functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined by

$$F(m) = \Phi(\langle \ell, m\rangle) = \Phi\biggl(\int \ell(x')m(\mathrm{d}x')\biggr),$$

the drift b is the negative intrinsic derivative:

$$b(m,x) = -D_mF(m,x) = -\nabla_x \frac{\delta F}{\delta m}(m,x).$$

We shall also impose the following conditions on a family of probability measures related to Φ and ℓ.

Assumption 7.12 (Uniform LSI)**.** The probability measures $(\hat{m}_y)_{y\in\mathcal{K}}$ on $\mathbb{R}^d$ determined by

$$\hat{m}_y(\mathrm{d}x) \propto \exp\bigl(-2\nabla\Phi(y) \cdot \ell(x)\bigr)\,\mathrm{d}x$$

are well defined and satisfy a uniform C_{LS}-logarithmic Sobolev inequality for some $C_{\mathrm{LS}} \geqslant 0$. That is to say, for all regular enough probability measure $m \in \mathcal{P}(\mathbb{R}^d)$ and all $y \in \mathcal{K}$, we have

$$\int \log\frac{\mathrm{d}m}{\mathrm{d}\hat{m}_y}\,\mathrm{d}m =: H(m|\hat{m}_y) \leqslant \frac{C_{\mathrm{LS}}}{4} I(m|\hat{m}_y) := \frac{C_{\mathrm{LS}}}{4}\int \biggl|\nabla\log\frac{\mathrm{d}m}{\mathrm{d}\hat{m}_y}\biggr|^2 \hat{m}_y,$$

where $\mathrm{d}m/\mathrm{d}\hat{m}_y$ is the Radon–Nikodým derivative between measures.

Remark 7.13. As mentioned in the introduction, the uniform log-Sobolev inequality is crucial to our method as it is used to obtain the entropy estimate in Proposition 7.25. This condition is often perceived as a strong one, but still it can be verified if for example

$$2\nabla\Phi(y) \cdot \ell(x) = U(x) + G(y,x)$$

for some strongly convex $U : \mathbb{R}^d \to \mathbb{R}$ and some bounded $G : \mathcal{K} \times \mathbb{R} \to \mathbb{R}$. Indeed, in this case $\hat{m}_y$ is a uniformly bounded perturbation of the strongly log-concave measure $e^{-U(x)}\,\mathrm{d}x \big/ \int e^{-U(x')}\,\mathrm{d}x'$, so it satisfies a uniform log-Sobolev inequality by the Bakry–Émery condition [11] and the Holley–Stroock perturbation [113]. We note that this condition has also been exploited recently to obtain long-time behaviors of mean field Langevin [56, 178] and its particle systems in Chapters 1 and 2.

Finally, we assume that the following quantitative bound on Φ and ℓ.

Assumption 7.14 (Bound)**.** The following quantity is finite:

$$\begin{aligned} M_2 &\coloneqq \sup_{x\in\mathbb{R}^d, y\in\mathbb{R}^D} \left|\nabla^2\Phi(y)^{1/2}\nabla\ell(x)\right|^2 \\ &= \sup_{x\in\mathbb{R}^d, y\in\mathbb{R}^D} \sup_{\substack{a\in\mathbb{R}^d\\ |a|=1}} a^\top \nabla\ell(x)^\top \nabla^2\Phi(y)\nabla\ell(x)a. \end{aligned}$$

Remark 7.15. Under the three assumptions above, if Φ is additionally convex, then there exists a unique invariant measure m_* of the McKean–Vlasov dynamics (7.1) by Proposition 1.32 and Corollary 1.39, and the convergence to the invariant measure is exponentially fast by Theorem 1.4. In fact, the convexity of Φ implies precisely the functional convexity of the mean field energy F considered in Chapter 1.

The main discovery of this paper characterizes the distance between the stationary measure P of the self-interacting dynamics (7.4) and $m_* \otimes \delta_{m_*}$.

Theorem 7.16. *Let Assumptions 7.11, 7.12 and 7.14 hold true. Suppose that $P = P^\lambda \in \mathcal{P}_4\big(\mathbb{R}^d \times \mathcal{P}_4(\mathbb{R}^d)\big)$ is a stationary measure of the self-interacting process (7.4) in the sense of Definition 7.9 that has finite fourth moment:*

$$\iint \left(|x|^4 + \int |x'|^4 m(\mathrm{d}x')\right) P(\mathrm{d}x\,\mathrm{d}m) < \infty.$$

Let (X, m) be a random variable distributed as P. Denote by ρ^1 and ρ^2 the probability distributions of the random variables X and $\langle \ell, m\rangle = \int \ell(x)m(\mathrm{d}x)$ respectively.

1. *Suppose Φ is convex. Denote in the case by m_* the unique invariant measure of the McKean–Vlasov dynamics (7.1). Define*

$$\begin{aligned} y_* &\coloneqq \langle \ell, m_*\rangle = \int \ell(x)m_*(\mathrm{d}x), \\ \lambda_0 &\coloneqq \frac{1}{48M_2C_{\mathrm{LS}}^2\Big(1 + 2M_2C_{\mathrm{LS}}\big(M_2^2C_{\mathrm{LS}}^2+1\big)^{1/2}\Big)}, \\ H &\coloneqq \frac{C_{\mathrm{LS}}(D + 24M_2C_{\mathrm{LS}}d)\lambda}{2 - 96M_2C_{\mathrm{LS}}^2\Big(1 + 2M_2C_{\mathrm{LS}}\big(M_2^2C_{\mathrm{LS}}^2+1\big)^{1/2}\Big)\lambda}. \end{aligned}$$

Then, for $\lambda \in (0, \lambda_0)$, we have

$$\begin{aligned} v(\rho^2) &\coloneqq \mathbb{E}\left[\iiint_0^1 \frac{\delta^2 F}{\delta m^2}\big((1-t)m + tm_*, x', x''\big)\,\mathrm{d}t\,(m - m_*)^{\otimes 2}(\mathrm{d}x'\,\mathrm{d}x'')\right] \\ &= \iint_0^1 (y - y_*)^\top \nabla^2\Phi\big((1-t)y + ty_*\big)(y - y_*)\,\mathrm{d}t\,\rho^2(\mathrm{d}y) \\ &\leqslant 4M_2C_{\mathrm{LS}}\big(M_2^2C_{\mathrm{LS}}^2 + 1\big)H, \end{aligned}$$

and

$$
\begin{aligned}
W_2^2(\rho^1, m_*) &\leqslant \Big(2C_{\mathrm{LS}} + 4M_2 C_{\mathrm{LS}}^2 \big(M_2^2 C_{\mathrm{LS}}^2 + 1\big)^{1/2}\Big) H,\\
\|\rho^1 - m_*\|_{\mathrm{TV}}^2 &\leqslant \Big(4 + 8M_2 C_{\mathrm{LS}} \big(M_2^2 C_{\mathrm{LS}}^2 + 1\big)^{1/2}\Big) H.
\end{aligned}
$$

2. *If in addition to the convexity of Φ,*

$$
M_1 \coloneqq \sup_{x\in\mathbb{R}^d, y\in\mathbb{R}^D} \ell(x)^\top \nabla^2\Phi(y)\ell(x) < \infty,
$$

then the three inequalities above hold also for all $\lambda \in (0,\infty)$, with H replaced by

$$
H' \coloneqq \frac{C_{\mathrm{LS}}}{2}(D + 2M_1)\lambda.
$$

3. *If Φ is concave, then for $\hat{y} \coloneqq \langle \ell, \hat{m}_y\rangle$, we have*

$$
-\int (\hat{y} - y)^\top \nabla^2\Phi(y)(\hat{y} - y)\rho^2(\mathrm{d}y) \leqslant \frac{M_2 C_{\mathrm{LS}}^2 D}{2}\lambda.
$$

The proof of Theorem 7.16 is postponed to Section 7.6.

Remark 7.17 (Dependence on the dimension D)*.* Readers may have observed that, in our framework, the value of the functional $F(m)$ may only depend on the D-dimensional vector $\int \ell(x)m(\mathrm{d}x)$, and this corresponds to "cylindrical functions" considered in [4, Definition 5.1.11]. Given a continuous functional F on $\mathcal{P}(\mathbb{R}^d)$, for every compact subset $\mathcal{S} \subset \mathbb{R}^d$, we can construct, according to the Stone–Weierstrass theorem, a sequence of functions $\ell_n : \mathbb{R}^d \to \mathbb{R}^{D_n}$ and $\Phi_n : \mathbb{R}^{D_n} \to \mathbb{R}$ such that the cylindrical functionals $F_n(m) = \Phi_n\big(\int \ell_n \,\mathrm{d}m\big)$ approach F in the uniform topology of $\mathcal{C}\big(\mathcal{P}(\mathcal{S})\big)$ (see [92, Lemma 2] for example). However, the dimension D_n may tend to infinity when $n \to \infty$. Since all the upper bounds in Theorem 7.16 depend on the dimension D linearly, our analysis and findings cannot be directly applied to more general functionals on $\mathcal{P}(\mathbb{R}^d)$.

Remark 7.18 (Meaning of M_1 and M_2)*.* The second-order functional derivative of F reads

$$
\frac{\delta^2 F}{\delta m^2}(m, x', x'') = \ell(x')^\top \nabla^2\Phi(y)\ell(x''),
$$

and in the case of convex Φ, satisfies the Cauchy–Schwarz inequality:

$$
\begin{aligned}
\left|\frac{\delta^2 F}{\delta m^2}(m, x', x'')\right| &= \big|\ell(x')^\top \nabla^2\Phi(y)\ell(x'')\big|\\
&\leqslant \big|\ell(x')^\top \nabla^2\Phi(y)\ell(x')\big|^{1/2} \big|\ell(x'')^\top \nabla^2\Phi(y)\ell(x'')\big|^{1/2}\\
&= \left|\frac{\delta^2 F}{\delta m^2}(m, x', x')\right|^{1/2} \left|\frac{\delta^2 F}{\delta m^2}(m, x'', x'')\right|^{1/2} \leqslant M_1.
\end{aligned}
$$

Similarly, the second-order intrinsic derivative satisfies

$$
\begin{aligned}
\big|D_m^2 F(m, x', x'')\big| &= \big|\nabla\ell(x')^\top \nabla^2\Phi(y)\nabla\ell(x'')\big|\\
&\leqslant \big|\nabla\ell(x')^\top \nabla^2\Phi(y)\nabla\ell(x')\big|^{1/2} \big|\nabla\ell(x'')^\top \nabla^2\Phi(y)\nabla\ell(x'')\big|^{1/2}\\
&= \big|D_m^2 F(m, x', x')\big|^{1/2} \big|D_m^2 F(m, x'', x'')\big|^{1/2} \leqslant M_2.
\end{aligned}
$$

Moreover, by taking $x' = x''$ in the inequalities above, we find that M_1 and M_2 are the respective suprema of the second-order flat and intrinsic derivatives of the functional F.

We illustrate in the following example that the order of λ when $\lambda \to 0$ for the variance of $\langle \ell, m\rangle$ under P in Theorem 7.16 (i.e., the first claim) is optimal.

Example 7.19 (Optimality of the order of λ). Consider the mean field functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ given by

$$F(m) = \frac{1}{2}\int |x|^2 m(\mathrm{d}x) + \frac{1}{2}\left|\int x m(\mathrm{d}x)\right|^2.$$

By taking $D = d+1$, $\ell(x) = (x, |x|^2/2)^\top$, $\Phi(y_0, y_1) = y_0^2/2 + y_1$, the mean field functional F is covered by the cylindrical setting (namely Assumption 7.11) of Theorem 7.16. Moreover, the function Φ is convex.

The corresponding gradient dynamics (7.1) is then characterized by the drift

$$b(m, x) = -x - \int x' m(\mathrm{d}x'),$$

and its unique invariant measure m_* is $\mathcal{N}(0, 1/2)$. We explicitly compute quantities related to the stationary measure P^λ of the self-interacting dynamics (7.4) in the following. The dynamics reads

$$\begin{aligned} \mathrm{d}X_t &= -X_t\,\mathrm{d}t - \int x' m_t(\mathrm{d}x')\,\mathrm{d}t + \mathrm{d}B_t, \\ \mathrm{d}m_t &= \lambda(\delta_{X_t} - m_t)\,\mathrm{d}t, \end{aligned}$$

and has a unique strong solution by Cauchy–Lipschitz arguments. Upon defining $Y_{0,t} = \int x' m_t(\mathrm{d}x')$, the process has the finite-dimensional projection:

$$\begin{aligned} \mathrm{d}X_t &= (-X_t - Y_{0,t})\,\mathrm{d}t + \mathrm{d}B_t, \\ \mathrm{d}Y_{0,t} &= \lambda(-Y_{0,t} + X_t)\,\mathrm{d}t. \end{aligned}$$

The finite-dimensional dynamics has a unique invariant measure that is a centered Gaussian with the following covariance structure:

$$\begin{aligned} \mathbb{E}[X \otimes X] &= \frac{\lambda+2}{4(\lambda+1)}\mathbb{1}_{d\times d}, \\ \mathbb{E}[Y_0 \otimes Y_0] = \mathbb{E}[X \otimes Y_0] &= \frac{\lambda}{4(\lambda+1)}\mathbb{1}_{d\times d}. \end{aligned}$$

Hence, the exact distances, or bounds thereof, read

$$\begin{aligned} \mathbb{E}\big[|Y_0|^2\big] &= \frac{d\lambda}{4(1+\lambda)}, \\ W_2^2\big(\mathrm{Law}(X), m_*\big) &= \frac{d}{2}\left(1 - \left(1 - \frac{\lambda}{2(1+\lambda)}\right)^{1/2}\right)^2, \\ \|\mathrm{Law}(X) - m_*\|_{\mathrm{TV}}^2 &\in \left[\frac{1}{10000}\frac{d\lambda^2}{4(1+\lambda)^2}, \frac{9}{4}\frac{d\lambda^2}{4(1+\lambda)^2}\right], \end{aligned}$$

where the last mutual bound is due to [71, Theorem 1.1].

Now we try to verify the assumptions of Theorem 7.16. By the Kolmogorov extension theorem, we can construct a stationary Markov process $(X_t, Y_{0,t})_{t\in\mathbb{R}}$, defined on the whole real line, such that $\mathrm{Law}(X_t, Y_{0,t}) = \mathrm{Law}(X, Y_0)$ for all $t \in \mathbb{R}$. Then, by defining

$$m_t = \lambda \int_{-\infty}^{t} e^{-\lambda(t-s)} \delta_{X_s}\, \mathrm{d}s,$$

we recover the solution $(X_t, m_t)_{t\in\mathbb{R}}$ to the original infinite-dimensional dynamics. By construction, $\mathrm{Law}(X_t, m_t)$ is stationary and has finite fourth moments. The rest of the assumptions of Theorem 7.16 can be satisfied with the constants $C_{\mathrm{LS}} = 1/2$ and $M_2 = 1$. Since Φ is convex, by the first claim of the theorem, we get

$$\begin{aligned}
\mathbb{E}\big[|Y_0|^2\big] &\leqslant \frac{5(13d+1)\lambda}{8 - 48(2+\sqrt{5})\lambda}, \\
W_2^2\big(\mathrm{Law}(X), m_*\big) &\leqslant \frac{(2+\sqrt{5})(13d+1)\lambda}{8 - 48(2+\sqrt{5})\lambda}, \\
\|\mathrm{Law}(X) - m_*\|_{\mathrm{TV}}^2 &\leqslant \frac{(2+\sqrt{5})(13d+1)\lambda}{2 - 12(2+\sqrt{5})\lambda},
\end{aligned}$$

for $\lambda < 1/6(2+\sqrt{5})$. So the results of Theorem 7.16 give the optimal order of λ when $\lambda \to 0$ for the variance of the Y variable, but possibly sub-optimal ones for the Wasserstein and total-variation distances in the X direction.

7.2.3 A class of dynamics

In this subsection, we present a class of dynamics to which both Theorems 7.4 and 7.16 are applicable. This class encompasses in particular the neural network example that will be discussed in the following Section 7.3.

Assumption 7.20. The drift functional writes

$$b(m,x) = -\nabla U(x) - \nabla \Phi_0\bigg(\int \ell_0(x')m(\mathrm{d}x')\bigg) \cdot \nabla \ell_0(x). \tag{7.6}$$

for some functions $U : \mathbb{R}^d \to \mathbb{R}$, $\Phi_0 : \mathbb{R}^D \to \mathbb{R}$, $\ell_0 : \mathbb{R}^d \to \mathbb{R}^D$ satisfying the following conditions:

- the function U is $\mathcal{C}^2$ continuous with bounded Hessian, i.e., $\|\nabla^2 U\|_{L^\infty} < \infty$, and its gradient admits a semi-monotonicity function $\kappa(x) = \kappa_0 - M/x$ for some $\kappa_0 > 0$ and $M \geqslant 0$.
- the probability measure $Z^{-1}\exp\big(-2U(x)\big)\,\mathrm{d}x$, with $Z = \int \exp\big(-2U(x)\big)\,\mathrm{d}x$, is well defined in $\mathcal{P}(\mathbb{R}^d)$, and satisfies a $C_{\mathrm{LS},0}$-logarithmic Sobolev inequality.
- the function Φ_0 is κ_{Φ_0}-strongly convex for some $\kappa_{\Phi_0} > 0$ and belongs to $\mathcal{C}^2(\mathbb{R}^D) \cap W^{2,\infty}(\mathbb{R}^D)$.
- the function ℓ_0 belongs to $\mathcal{C}^1(\mathbb{R}^d;\mathbb{R}^D) \cap W^{1,\infty}(\mathbb{R}^d;\mathbb{R}^D)$.

Proposition 7.21. *Under Assumption 7.20, there exists a unique stationary measure $P \in \mathcal{P}_1\big(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\big)$ to the self-interacting dynamics (7.4). Moreover, there exists $C > 0$, independent of λ, such that for (X, m) distributed as P,*

$$\mathbb{E}[|\langle \ell_0, m - m_*\rangle|^2] + W_2^2\big(\mathrm{Law}(X), m_*\big) + \|\mathrm{Law}(X) - m_*\|_{\mathrm{TV}}^2 \leqslant C\lambda,$$

where m_* is the unique invariant measure to the McKean–Vlasov process.

Proof of Proposition 7.21. We first verify the conditions of Theorem 7.4 to establish the existence and uniqueness of the stationary measure. As the drift b has derivative

$$\frac{\delta b}{\delta m}(m, x, x') = -\nabla^2\Phi_0\bigg(\int \ell_0(x'')m(\mathrm{d}x'')\bigg) \cdot \nabla\ell_0(x) \cdot \ell_0(x'),$$

we have

$$\bigg\|\frac{\delta b}{\delta m}(m, x, \cdot)\bigg\|_{W^{1,\infty}} \leqslant \|\nabla^2\Phi_0\|_{L^\infty}\|\nabla\ell_0\|_{L^\infty}\|\ell_0\|_{W^{1,\infty}}$$

for every $m \in \mathcal{P}(\mathbb{R}^d)$ and every $x \in \mathbb{R}^d$. Then the dynamics falls into the class considered in Example 7.8. The conditions of Theorem 7.4 are satisfied. Applying Corollary 7.10 gives the existence and the uniqueness of the stationary measure $P \in \mathcal{P}_1\big(\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)\big)$.

We proceed to verify the conditions of Theorem 7.16. Introduce the functions $\Phi : \mathbb{R}^{D+1} \to \mathbb{R}$, $\ell : \mathbb{R}^d \to \mathbb{R}^{D+1}$ defined respectively by

$$\begin{aligned} \Phi(y_0, y_1) &= \Phi_0(y_0) + y_1, && \text{for } y_0 \in \mathbb{R}^D \text{ and } y_1 \in \mathbb{R},\\ \ell(x) &= \big(\ell_0(x), U(x)\big), && \text{for } x \in \mathbb{R}^d. \end{aligned}$$

Here the range set $\mathcal{K}$ of the mapping ℓ is taken as the whole space $\mathbb{R}^D$. In this way, the drift b reads

$$b(m, x) = -\nabla\Phi\bigg(\int \ell(x')m(\mathrm{d}x')\bigg) \cdot \nabla\ell(x)$$

so Assumption 7.11 is satisfied. Now we show the stationary measure P of the dynamics (7.4) has finite fourth moment. Consider the functional

$$E(x, m) = \kappa_0^{-1}|x|^4 + \int |x'|^4 m(\mathrm{d}x').$$

Along the self-interacting dynamics (7.4), we have, by Itō's formula,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[E(X_t, m_t)] &= 4\kappa_0^{-1}\,\mathbb{E}\big[b(m_t, X_t) \cdot X_t|X_t|^2\big] + (2d + 4)\kappa_0^{-1}\,\mathbb{E}\big[|X_t|^2\big]\\ &\quad - \lambda\,\mathbb{E}\bigg[\int |x'|^4 m'(\mathrm{d}x')\bigg] + \lambda\,\mathbb{E}\big[|X_t|^4\big]. \end{aligned}$$

As the vector field $x \mapsto b(m, x)$ has weak monotonicity function $\kappa_b(x) = \kappa_0 - M_b/x$, we have

$$b(m, x) \cdot x \leqslant -\frac{\kappa_0}{2}|x|^2 + C_2$$

for every $(m, x) \in \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d$, for some $C_2 \geqslant 0$. The functional E verifies the Lyapunov condition as well:

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[E(X_t, m_t)] \leqslant -c_3\,\mathbb{E}[E(X_t, m_t)] + C_3$$

for some $c_3 > 0$, $C_3 \geqslant 0$. In consequence, the stationary measure P of finite first moment must have finite fourth moment. Since

$$\hat{m}_y(\mathrm{d}x) \propto \exp\bigl(-2\nabla\Phi(y)\cdot\ell(x)\bigr)\,\mathrm{d}x = \exp\bigl(-2\nabla\Phi_0(y)\cdot\ell_0(x)\bigr)\exp\bigl(-2U(x)\bigr)\,\mathrm{d}x$$

with $\sup_{x,y}|\nabla\Phi_0(y)\cdot\ell_0(x)| \leqslant \|\nabla\Phi_0\|_{L^\infty}\|\ell_0\|_{L^\infty}$, by the Holley–Stroock perturbation argument [113], we know that the measure $\hat{m}_y$ verifies a uniform C_{LS}-LSI with

$$C_{\mathrm{LS}} = C_{\mathrm{LS},0}\exp(4\|\nabla\Phi_0\|_{L^\infty}\|\ell_0\|_{L^\infty}),$$

so Assumption 7.12 is satisfied. The constants M_1, M_2 in Theorem 7.16 and Assumption 7.14 are finite as

$$\begin{aligned}
|\ell(x)^\top\nabla^2\Phi(y)\ell(x)| &= |\ell_0(x)^\top\nabla^2\Phi_0(y)\ell_0(x)| \leqslant \|\nabla^2\Phi_0\|_{L^\infty}\|\ell_0\|_{L^\infty}^2,\\
|\nabla\ell(x)\nabla^2\Phi(y)\nabla\ell(x)| &= |\nabla\ell_0(x)^\top\nabla^2\Phi_0(y)\nabla\ell_0(x)| \leqslant \|\nabla^2\Phi_0\|_{L^\infty}^2\|\nabla\ell_0\|_{L^\infty}.
\end{aligned}$$

All the conditions of Theorem 7.16 are satisfied. Since for all $y \in \mathbb{R}^D$,

$$\begin{aligned}
\kappa_{\Phi_0}|\langle\ell_0, m - m_*\rangle|^2 &\leqslant \langle\ell_0, m - m_*\rangle^\perp\nabla^2\Phi_0(y)\langle\ell_0, m - m_*\rangle\\
&= \langle\ell, m - m_*\rangle^\perp\nabla^2\Phi(y)\langle\ell, m - m_*\rangle,
\end{aligned}$$

the claim of the proposition follows from the second case stated in the theorem. □

Remark 7.22 (Open question). Applying further the convergence result of Theorems 7.4, we can obtain a bound on the difference between the marginal distribution of the non-stationary self-interacting diffusion (7.4) and the invariant measure of the McKean–Vlasov dynamics (7.1). Specifically, let (X_t, m_t) denote the self-interacting process (7.4). Theorem 7.4 yields

$$W_1\bigl(\mathrm{Law}(X_t, m_t), \mathrm{Law}(X, m)\bigr) \leqslant Ce^{-ct},$$

where C, c are the contractivity constants in the theorem. Let $\varphi : \mathbb{R}^K \to \mathbb{R}$ be a 1-Lipschitz test function. Combining this with Proposition 7.21, we can bound the following L^1 simulation error:

$$\mathbb{E}[|\langle\varphi\circ\ell, m_t - m_*\rangle|] = \mathbb{E}[|\langle\varphi\circ\ell, m_t - m + m - m_*\rangle|] = O(e^{-ct} + \sqrt{\lambda}).$$

As noted in Remark 7.6, the contraction rate c depends on $1/\lambda$ exponentially, rendering the above error bound impractical for applications.

This naturally raises the question of whether the method and results of Theorem 7.16, which address the static case (i.e., the comparison between $\mathrm{Law}(X, m)$ and $m_*\otimes\delta_{m_*}$), can be extended to the dynamical setting (comparing $\mathrm{Law}(X_t, m_t)$ and $m_*\otimes\delta_{m_*}$). Unfortunately, we are currently unable to establish the entropy estimate in Section 7.6.2 for the parabolic problem, so our approach does not yet apply in this context. We leave this as an open problem for future research.

7.3 Numerical application to training two-layer neural networks

Let us recall the structure of two-layer neural networks and introduce the mean field model for it. Consider an *activation function* $\varphi : \mathbb{R} \to \mathbb{R}$ satisfying

$$\begin{gathered}
\varphi \text{ is continuous and non-decreasing},\\
\lim_{\theta\to-\infty}\varphi(\theta) = 0, \qquad \lim_{\theta\to+\infty}\varphi(\theta) = 1.
\end{gathered}$$

Define $S = \mathbb{R} \times \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}$, where the *neurons* take values. For each neuron $x = (c, a, b) \in S$ we define the *feature map*:

$$\mathbb{R}^{d_{d_{\text{in}}}} \ni z \mapsto f(x; z) \coloneqq \tau(c)\varphi(a \cdot z + b) \in \mathbb{R},$$

where $\tau : \mathbb{R} \to [-L, L]$ is a *truncation function* with the *truncation threshold* $L \in (0, +\infty)$. The two-layer neural network is nothing but the averaged feature map parameterized by N neurons $x^1, \ldots, x^N \in S$:

$$\mathbb{R}^{d_{\text{in}}} \ni z \mapsto f^N(x^1, \ldots, x^N; z) = \frac{1}{N} \sum_{i=1}^{N} f(x^i; z) \in \mathbb{R}. \tag{7.7}$$

The training of neural network aims to minimize the distance between the averaged output (7.7) for K data points $(z_1, \ldots, z_K)$ and their labels $(y_1, \ldots, y_K)$, that is, the objective function of the minimization reads

$$F^N_{\text{NNet}}(x^1, \ldots, x^N) = \frac{N}{2K} \sum_{i=1}^{K} \big| y_i - f^N(x^1, \ldots, x^N; z_i) \big|^2. \tag{7.8}$$

This objective function is of high dimension and non-convex, and this difficulty motivates the recent studies, see for example [163, 57, 117] among others, to lift the finite-dimensional optimization (7.8) to the space of probability measures and to consider the following mean field optimization:

$$\inf_{m \in \mathcal{P}_2(S)} F_{\text{NNet}}(m), \quad \text{where } F_{\text{NNet}}(m) \coloneqq \frac{1}{2K} \sum_{i=1}^{K} \big| y_i - \mathbb{E}^{X \sim m}[f(X; z_i)] \big|^2.$$

The mean field loss functional F_{NNet} is apparently convex. Given that the activation and truncation functions φ, τ have bounded derivatives of up to fourth order, it has been proved in Proposition 1.34 that the minimum of the entropy-regularized mean field optimization problem

$$\inf_{m \in \mathcal{P}_2(S)} F_{\text{NNet}}(m) + \frac{\sigma^2}{2} H\big(m \big| \mathcal{N}(0, \gamma^{-1})\big)$$

can be attained by the invariant measure of the mean field Langevin dynamics:

$$\mathrm{d}X_t = -D_m F\big(\mathrm{Law}(X_t), X_t\big)\, \mathrm{d}t + \sigma\, \mathrm{d}W_t, \tag{7.9}$$

where the mean field potential functional reads

$$F(m) \coloneqq F_{\text{NNet}}(m) + \frac{\sigma^2 \gamma}{2} \int |x|^2 m(\mathrm{d}x).$$

By defining

$$\ell^0(x) \coloneqq |x|^2, \qquad \ell^i(x) \coloneqq f(x, z_i) - y_i, \quad \text{for } i = 1, \ldots, K,$$
$$R^{K+1} \ni \theta = (\theta_0, \theta_1, \ldots, \theta_K) \mapsto \Phi(\theta) \coloneqq \frac{\sigma^2 \gamma}{2} \theta_0 + \frac{1}{2K} \sum_{i=1}^{K} |\theta_i|^2,$$

we note that the mean field potential functional is of the form:

$$F(m) = \Phi(\langle \ell, m \rangle), \quad \text{where } \ell \coloneqq (\ell^0, \ell^1, \ldots, \ell^K),$$

as in the gradient case investigated in Sections 7.2.2.

In order to simulate the invariant measure of the mean field Langevin dynamics (7.9), one usually turns to the corresponding N-particle system:

$$\mathrm{d}X_t^j = \big(-\nabla_j F^N_{\mathrm{NNet}}(X_t^1,\dots,X_t^N) - \sigma^2\gamma X_t^j\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t^j, \quad \text{for } j = 1,\dots,N \tag{7.10}$$

The uniform-in-time propagation of chaos results obtained in [215] and Chapter 1 ensure that the marginal distributions $\mathrm{Law}(X_t^1,\dots,X_t^N)$ of the N-particle system are close to those of the mean field Langevin dynamics uniformly on the whole time horizon, and can efficiently approximate mean field invariant measure provided that t and N are both large enough.

Note that the N-particle system (7.10) is nothing but a regularized and noised gradient descent flow for N neurons. In contrast, the self-interacting diffusion

$$\begin{aligned}
\mathrm{d}X_t &= -\sum_{i=0}^{K} \nabla_i \Phi\big(Y_t^0, Y_t^1,\dots,Y_t^K\big)\nabla \ell^i(X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t \\
&= -\bigg(\frac{1}{K}\sum_{k=1}^{K} Y_t^k \nabla f(X_t, z_k) + \sigma^2\gamma X_t\bigg)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \\
\mathrm{d}Y_t^i &= \lambda\big(\ell^i(X_t) - Y_t^i\big)\,\mathrm{d}t, \quad \text{for } i = 1,\dots,K,
\end{aligned} \tag{7.11}$$

introduces an innovative alternative algorithm for training two-layer neural networks, in which the algorithm iterations impact only a single neuron.

Setup. We aim to train a neural network to approximate the non-linear elementary function $\mathbb{R}^2 \ni z = (z_1, z_2) \mapsto g(z) := \sin 2\pi z_1 + \cos 2\pi z_2$. Note that in this numerical example $d_{\mathrm{in}} = 2$ and therefore $S = \mathbb{R}^{1+2+1}$. We draw K points $\{z_i\}_{k=1}^K$ sampled according to the uniform distribution on $[0,1]^2$ and compute the corresponding labels $y_k = g(z_k)$ to form our training data $\{z_k, y_k\}_{k=1}^K$. We fix the truncation function τ by $\tau(\theta) = (\theta \wedge 30) \vee -30$ and the sigmoid activation function φ by $\varphi(\theta) = 1/\big(1+\exp(-\theta)\big)$. The Brownian noise has volatility σ such that $\sigma^2/2 = 0.05$, and the regularization constant γ is fixed at $\gamma = 0.0025$ in the experiment. The initial value $X_0 = (C_0, A_0, B_0)$, taking values in $S = \mathbb{R}^{1+2+1}$, is sampled from the normal distribution $m_0 = \mathcal{N}(0, 10^2 \times \mathrm{Id}_{4\times 4})$ in four dimensions. To observe the impact of the self-interacting coefficient λ, we run the simulation of (7.11) for different λ equal to 4^{-p} for $p = 4,\dots,8$. Furthermore, to compare with these results with fixed λ, we choose the non-increasing piecewise constant function λ_{a} such that $\lambda_{\mathrm{a}}(t) = 4^{-i}$ on successive intervals of length $\delta T_i = 4^i$, for $i = 2,\dots,11$, and train the neural network along the annealing scheme:

$$\begin{aligned}
\mathrm{d}X_t &= -\bigg(\frac{1}{K}\sum_{k=1}^{K} Y_t^k \nabla f(X_t, z_k) + \sigma^2\gamma X_t\bigg)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \\
\mathrm{d}Y_t^i &= \lambda_{\mathrm{a}}(t)\big(\ell^i(X_t) - Y_t^i\big)\,\mathrm{d}t, \quad \text{for } i = 1,\dots,K.
\end{aligned} \tag{7.12}$$

We simulate both the constant and dynamic-λ self-interacting diffusions (7.11), (7.12) by the Euler scheme, as described in Appendix D.2, on a long time horizon till the terminal time $T = 10^6$, with the discrete time step $\delta t = 0.5$.

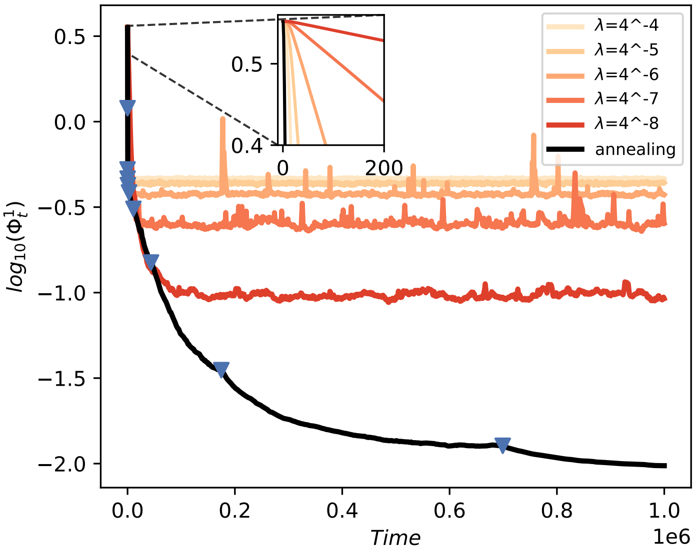


Figure 7.1: Averaged over 100 repetitions losses for fixed values of λ and for discrete annealing.

Results and discussions. We repeat the simulations mentioned above for fixed λ's and dynamic λ_{a} all for 100 times and compute the averaged loss $\Phi(Y_t)$ at each time t and plot its evolution in Figure 7.1. On the annealing scheme curve, we include triangles to visually indicate the points at which there are changes on the values of λ_{a}. We observe that the value of $\Phi(Y_t)$ first decays exponentially, and the speed of initial decay depending on the value of λ. More precisely, the bigger is the value of λ, the faster is the initial decay. In particular, this remains true for the annealed process as it starts from a bigger value $\lambda_{\mathrm{a}}(0)$. We notice that such phenomenon is in line with our theoretical results in Theorem 7.4. In the long run, when fixing λ, the value of $\Phi(Y_t)$ stabilizes at a level sensible with respect to the value of λ. We notice that the smaller is the value of λ, the better is the final performance. These facts are in line with our discovery in Theorems 7.4 and 7.16. Finally, the training exhibits the best performance when implementing the discrete annealing. The loss $\Phi(Y_t)$ continues to decrease as the value of $\lambda_{\mathrm{a}}(t)$ diminishes.

7.4 A Curie–Weiss model

In this section, we present and study a simple Curie–Weiss model, i.e., a mean field model of ferromagnets, which has possibly multiple invariant measures. In particular, we show that in this case, the last claim of Theorem 7.16 provides informations on the concentration of the self-interacting stationary measure.

Let $\ell_0 : \mathbb{R} \to \mathbb{R}$ be a smooth, odd, increasing function in $\mathcal{C}^1 \cap W^{1,\infty}$. For a probability measure $m \in \mathcal{P}(\mathbb{R})$, consider the mean field energy

$$F(m) = -\frac{1}{2}\Big(\int \ell_0(x)m(\mathrm{d}x)\Big)^2 + \frac{1}{2}\int x^2 m(\mathrm{d}x).$$

By setting

$$\ell(x) := \big(\ell_0(x), |x|^2/2\big)^\top,$$
$$\Phi(y_0, y_1) := -\frac{1}{2}|y_0|^2 + y_1,$$

we have $F(m) = \Phi(\langle \ell, m\rangle)$. So we are in the cylindrical setting of Assumption 7.11 with the range set of ℓ being defined by

$$\mathcal{K} := [-\|\ell_0\|_{L^\infty}, \|\ell_0\|_{L^\infty}] \times \mathbb{R}.$$

Moreover, as the corresponding intrinsic derivative, or the drift, writes

$$b(m, x) = -D_m F(m, x) = \langle \ell_0, m\rangle \ell_0'(x) - x,$$

we can verify Assumption 7.3 with the modulus of continuity

$$\omega(r) = \sup_{x,x' \in \mathbb{R} : |x - x'| \leqslant r} |\ell_0(x) - \ell_0(x')|.$$

This implies, by Corollary 7.10, that the self-interacting process has a unique invariant measure, which we denote by $\rho^\lambda = \rho$. Arguing as in the proof of Proposition 7.21, we can verify the conditions of Theorem 7.16 for the choice of Φ, ℓ and ρ above. Especially, the probability measures

$$\hat{m}_{(y_0,y_1)}(\mathrm{d}x) \propto \exp\biggl(-\frac{1}{2}|x|^2 + y_0\ell_0(x)\biggr)\,\mathrm{d}x$$

satisfy a uniform LSI thanks to the boundedness of y_0 and the Holley–Stroock perturbation lemma.

Before applying Theorem 7.16, we first give a characterization of the invariant measure for the McKean–Vlasov dynamics (7.1). We discuss a special case and then move to general discussions.

1. If $\|\ell'\|_{L^\infty} < 1$, then by the result of [224], we already know that the McKean–Vlasov dynamics (7.1) has a unique invariant measure, which is the centered Gaussian of variance $1/2$, i.e., $\mathcal{N}(0, 1/2)$. This case corresponds to the weak interaction regime studied in [75].

2. In the general case where $\|\ell'\|_{L^\infty}$ is not restricted, the invariant measures of (7.1) correspond to fixed points of the one-dimensional mapping

   $$\mathbb{R} \ni y_0 \mapsto \Pi_0(y_0) = \hat{y}_0 \in \mathbb{R}$$

   given by

   $$\Pi_0(y_0) = \frac{\int \ell_0(x)\exp\bigl(2y_0\ell_0(x) - |x|^2\bigr)\,\mathrm{d}x}{\int \exp\bigl(2y_0\ell_0(x) - |x|^2\bigr)\,\mathrm{d}x}$$

   That is to say, if y_0 satisfies $\Pi_0(y_0) = y_0$, then the probability measure proportional to $\exp\bigl(2y_0\ell_0(x) - |x|^2\bigr)\,\mathrm{d}x$ is invariant to (7.1); and vice versa. Due to the oddness of ℓ_0, the mapping Π_0 is odd. In particular, $\Pi_0(0) = 0$ and we know that $\mathcal{N}(0, 1/2)$ is always invariant.

   If $\Pi_0'(0) > 1$, then by the fact that $\|\Pi_0\|_{L^\infty} \leqslant \|\ell_0\|_{L^\infty} < \infty$ and the intermediate value theorem, there exists $y_0 > 0$ such that $\Pi_0(y_0) = y_0$

and $\Pi_0(-y_0) = -y_0$. That is to say there exists at least three invariant measures, two of which are, in physicists' language, "symmetry-breaking" phases, and the centered Gaussian measure is the "symmetry-preserving" phase. Moreover, the centered measure corresponding to $y_0 = 0$ should be unstable as it is a local maximum point for the free energy landscape $y_0 \mapsto \frac{1}{2}y_0^2 - \frac{1}{2}\log\int\exp\bigl(2y_0\ell_0(x) - |x|^2\bigr)\,\mathrm{d}x$.

We now turn to the study of the stationary measure ρ^λ of the self-interacting process (7.4). Since $\Phi(y) = -|y_0|^2/2 + y_1$ is linear in its second coordinate, the last claim of Theorem 7.16 implies that

$$\int |\hat{y}_0 - y_0|^2 \rho^2(\mathrm{d}y) = -\int (\hat{y} - y)^\top \nabla^2\Phi(y)(\hat{y} - y)\rho^2(\mathrm{d}y) \leqslant C\lambda,$$

where C is a constant depending only on ℓ_0 and as we recall, y_0 is the first coordinate of y. In other words, the stationary measure ρ^λ solves the self-consistency equation

$$\hat{y}_0 = y_0$$

up to an error of order $O(\sqrt{\lambda})$. Denote the set of fixed points of $y_0 \mapsto \Pi_0(y_0) = \hat{y}_0$ by S. If the set S is finite, and if for each $s \in S$ we have

$$\Pi_0'(s) \neq 1,$$

then there exists $c > 0$ such that for all $y_0 \in \mathbb{R}$,

$$|\hat{y}_0 - y_0| \geqslant c\min\bigl(d(y_0, S), 1\bigr),$$

where $d(\cdot, S)$ indicates the distance to the set S. In this case, we have

$$\int \min\bigl(d(y_0, S), 1\bigr)^2 \rho^2(\mathrm{d}y) = O(\lambda).$$

That is to say, for small λ, the random variable $Y = (Y_0, Y_1)$, distributed as the second component of the stationary measure ρ^λ, has Y_0 concentrated near the solutions to the self-consistency equation, which correspond to invariant measures of the McKean–Vlasov dynamics. However, the last claim of Theorem 7.16 is not sufficient to show that Y_0 is only concentrated around, or in a way "selects", the physically stable phases that are minimizers of a free energy functional. We refer readers to [17] for qualitative results on such selection mechanism.

7.5 Proof of Theorem 7.4

We first note that as the metric space $\bigl(\mathcal{P}(\mathbb{R}^d), W_\omega\bigr)$ is separable, we do not have measurability issues. We refer readers to [149, Chapter 2] for details.

Recall that the self-interacting dynamics (7.4) writes

$$\begin{aligned}\mathrm{d}X_t &= b(m_t, X_t)\,\mathrm{d}t + \mathrm{d}B_t,\\ \mathrm{d}m_t &= \lambda(-m_t + \delta_{X_t})\,\mathrm{d}t\end{aligned}$$

and similarly for the other dynamics (X', m') driven by another Brownian motion B'. Fix $n \in \mathbb{N}$. Let $\mathrm{rc} : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ be a Lipschitz continuous function such that $\mathrm{sc} := \sqrt{1 - \mathrm{rc}^2}$ is also Lipschitz continuous and

$$\mathrm{rc}(x, x') = \begin{cases} 1 & \text{if } |x - x'| \geqslant 2n^{-1},\\ 0 & \text{if } |x - x'| \leqslant n^{-1}.\end{cases}$$

We couple the two dynamics (X, m), (X', m') by

$$\begin{aligned} \mathrm{d}B_t &= \mathrm{rc}(X_t, X'_t)\,\mathrm{d}B^{\mathrm{rc}}_t + \mathrm{sc}(X_t, X'_t)\,\mathrm{d}B^{\mathrm{sc}}_t, \\ \mathrm{d}B'_t &= \mathrm{rc}(X_t, X'_t)\big(1 - 2e_te_t^\top\big)\,\mathrm{d}B^{\mathrm{rc}}_t + \mathrm{sc}(X_t, X'_t)\,\mathrm{d}B^{\mathrm{sc}}_t, \end{aligned}$$

where B^{rc}, B^{sc} are independent Brownian motions and e_t is the d-dimensional vector defined by

$$e_t = \begin{cases} \frac{X_t - X'_t}{|X_t - X'_t|} & \text{if } X_t \neq X'_t, \\ (1, 0, \dots, 0)^\top & \text{otherwise.} \end{cases}$$

Subtracting the dynamical equations of X, X' and denoting $\delta X = X - X'$, we obtain

$$\mathrm{d}\delta X_t = \big(b(m_t, X_t) - b(m'_t, X'_t)\big)\,\mathrm{d}t + 2\,\mathrm{rc}(X_t, X'_t)e_t\,\mathrm{d}W_t,$$

where $W_t \coloneqq \int_0^t e_t^\top\,\mathrm{d}B^{\mathrm{rc}}_t$ and is a one-dimensional Brownian motion by Lévy's characterization. The absolute value of the semimartingale δX_t admits the decomposition $\mathrm{d}|\delta X_t| = \mathrm{d}A^{|\delta X|}_t + \mathrm{d}M^{|\delta X|}_t$ with

$$\begin{aligned} \mathrm{d}A^{|\delta X|}_t &\leqslant -|\delta X_t|\kappa(|\delta X_t|)\,\mathrm{d}t + LW_\omega(m_t, m'_t)\,\mathrm{d}t, \\ \mathrm{d}M^{|\delta X|}_t &= 2\,\mathrm{rc}(X_t, X'_t)\,\mathrm{d}W_t. \end{aligned}$$

For the m variable, we have

$$\mathrm{d}(m_t - m'_t) = -\lambda(m_t - m'_t)\,\mathrm{d}t + \lambda(\delta_{X_t} - \delta_{X'_t})\,\mathrm{d}t.$$

Thus, by considering the (random) W_ω-optimal coupling at each t, we get

$$\mathrm{d}W_\omega(m_t, m'_t) \leqslant -\lambda W_\omega(m_t, m'_t)\,\mathrm{d}t + \lambda\omega(|\delta X_t|)\,\mathrm{d}t.$$

Hence the process

$$r_t = |\delta X_t| + \frac{2L}{\lambda}W_\omega(m_t, m'_t)$$

admits the decomposition $\mathrm{d}r_t = \mathrm{d}A^r_t + \mathrm{d}M^r_t$ with

$$\begin{aligned} \mathrm{d}A^r_t &\leqslant \big(-|\delta X_t|\kappa(|\delta X_t|) + 2L\omega(|\delta X_t|) - LW_\omega(m_t, m'_t)\big)\,\mathrm{d}t, \\ \mathrm{d}M^r_t &= 2\,\mathrm{rc}(X_t, X'_t)\,\mathrm{d}W_t. \end{aligned}$$

Let $f : [0, \infty) \to [0, \infty)$ be a $\mathcal{C}^2$ function to be determined in the following. We define $\rho_t = f(r_t)$. The process ρ_t admits the decomposition $\mathrm{d}\rho_t = \mathrm{d}A^\rho_t + \mathrm{d}M^\rho_t$ with

$$\begin{aligned} \mathrm{d}A^\rho_t &\leqslant \big(-|\delta X_t|\kappa(|\delta X_t|) + 2L\omega(|\delta X_t|) - LW_\omega(m_t, m'_t)\big)f'_-(r_t)\,\mathrm{d}t \\ &\quad + 2\,\mathrm{rc}(X_t, X'_t)^2 f''(r_t)\,\mathrm{d}t \\ &\leqslant -r_t\tilde{K}(r_t)f'_-(r_t)\,\mathrm{d}t + 2\,\mathrm{rc}(X_t, X'_t)^2 f''(r_t)\,\mathrm{d}t \end{aligned}$$

for the function $\tilde{K} : (0, \infty) \to \mathbb{R}$ defined by

$$\begin{aligned} \tilde{K}(r) &\coloneqq \inf_{\substack{x + 2L\lambda^{-1}y = r \\ x, y > 0}} \frac{x\kappa(x) - 2L\omega(x) + Ly}{r} \\ &\geqslant \inf_{\substack{x + 2L\lambda^{-1}y = r \\ x, y > 0}} \frac{\kappa_0 x + Ly - M_b - 2LM_\omega}{r} \\ &\geqslant \min\left(\kappa_0, \frac{\lambda}{2}\right) - \frac{M_b + 2LM_\omega}{r} \\ &\eqqcolon K_0 - \frac{M}{r} \eqqcolon K(r). \end{aligned}$$

Thus, we have shown

$$\mathrm{d}A_t^\rho \leqslant -r_t K(r_t) f'_-(r_t)\,\mathrm{d}t + 2\,\mathrm{rc}(X_t, X'_t)^2 f''(r_t)\,\mathrm{d}t.$$

Following Du et al. [76], we choose the function $f : [0,\infty) \to [0,\infty)$ by requiring

$$f'(r) = \frac{1}{2}\int_r^\infty s\exp\biggl(-\frac{1}{2}\int_r^s \tau K(\tau)\,\mathrm{d}\tau\biggr)\,\mathrm{d}s.$$

and $f(0) = 0$. The function f is indeed $\mathcal{C}^2$ continuous, and, according to [76, Lemma 5.1], it is also concave and satisfies

$$2f''(r) - rK(r)f'(r) + r = 0$$

and

$$\frac{1}{K_0} \leqslant f'(r),\ \frac{f(r)}{r} \leqslant f'(0)$$

for all $r > 0$. Plugging in the expression for K, we obtain

$$\begin{aligned}
f'(0) &= \frac{1}{2}\int_0^\infty s\exp\biggl(-\frac{K_0 s^2}{4} + \frac{Ms}{2}\biggr)\,\mathrm{d}s \\
&= \frac{1}{2}\exp\biggl(\frac{M^2}{4K_0}\biggr)\int_0^\infty s\exp\biggl(-\frac{K_0(s-\frac{M}{K_0})^2}{4}\biggr)\,\mathrm{d}s \\
&= \frac{1}{2}\exp\biggl(\frac{M^2}{4K_0}\biggr)\int_{-M/\sqrt{2K_0}}^\infty \biggl(\frac{2t}{K_0} + M\frac{2^{1/2}}{K_0^{3/2}}\biggr)\exp\biggl(-\frac{t^2}{2}\biggr)\,\mathrm{d}t \\
&\leqslant \frac{1}{2}\exp\biggl(\frac{M^2}{4K_0}\biggr)\biggl(\frac{2}{K_0}\exp\biggl(-\frac{M^2}{4K_0}\biggr) + 2M\frac{\pi^{1/2}}{K_0^{3/2}}\biggr) \\
&\leqslant \frac{1}{K_0} + \frac{2M}{K_0^{3/2}}\exp\biggl(\frac{M^2}{4K_0}\biggr).
\end{aligned}$$

For $|\delta X_t| \geqslant 2n^{-1}$, we have $\mathrm{rc}_t(X_t, X'_t) = 1$, and by the previous construction,

$$\mathrm{d}f(r_t) = \mathrm{d}A_t^\rho + \mathrm{d}M_t^\rho \leqslant -r_t\,\mathrm{d}t + \mathrm{d}M_t^\rho \leqslant -\frac{f(r_t)}{K_0^{-1} + 2MK_0^{-3/2}\exp(M^2/4K_0)}\,\mathrm{d}t + \mathrm{d}M_t^\rho.$$

For $|\delta X_t| < 2n^{-1}$, we proceed differently. Let ω_b denote the uniform modulus of equicontinuity of the vector fields $x \mapsto b(m,x)$. The absolute continuous part of r_t satisfies

$$\begin{aligned}
\mathrm{d}A_t^r &\leqslant \bigl(\omega_b(2n^{-1}) + 2L\omega(2n^{-1}) - LW_\omega(m_t, m'_t)\bigr)\,\mathrm{d}t \\
&= -LW_\omega(m_t, m'_t)\,\mathrm{d}t + o(1)\,\mathrm{d}t \\
&\leqslant -\frac{\lambda}{2}r_t\,\mathrm{d}t + o(1)\,\mathrm{d}t,
\end{aligned}$$

where $o(1)$ denotes a term that tends to 0 when $n \to \infty$.

Now we combine the two cases. Define a sequence of functions by

$$f_n(r) = \begin{cases} f(r), & \text{if } r \geqslant 2n^{-1}, \\ \frac{f(2n^{-1})}{2n^{-1}}r, & \text{if } r < 2n^{-1}. \end{cases}$$

Each function f_n is concave and satisfies, by the arguments above,

$$\mathrm{d}\,\mathbb{E}[f_n(r_t)] \leqslant -c\,\mathbb{E}[f_n(r_t)]\,\mathrm{d}t + o(1)\,\mathrm{d}t,$$

where

$$\begin{aligned} c &\coloneqq \min\biggl(\frac{1}{K_0^{-1} + 2MK_0^{-3/2}\exp(M^2/4K_0)}, \frac{\lambda}{2}\biggr) \\ &= \frac{1}{K_0^{-1} + 2MK_0^{-3/2}\exp(M^2/4K_0)}. \end{aligned}$$

Applying Grönwall's lemma, we obtain

$$\mathbb{E}[f_n(r_t)] \leqslant e^{-ct}\,\mathbb{E}[f_n(r_0)] + o(1).$$

Since $\lim_{n\to\infty} \mathbb{E}[f_n(r_t)] = \mathbb{E}[f(r_t)]$ by dominated convergence, taking the limit $n \to \infty$ completes the proof. □

7.6 Proof of Theorem 7.16

This section consists of four subsections. We show a series of intermediate, and sometimes technical, lemmas and propositions in the first three subsections before proving the theorem in the last subsection.

7.6.1 Elliptic equation for the stationary measure

We first note that the stationary measure P solves a partial differential equation in the following weak sense.

Proposition 7.23. *Let $\mathcal{C}_{\mathrm{b}}^{1,2}$ denote the class of functionals $\phi : \mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}$ with continuous spatial derivatives $\nabla_x\phi$, $\nabla_x^2\phi$ and a bounded linear functional derivative $\frac{\delta\phi}{\delta m} : \mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$ satisfying*

$$\begin{aligned} &\forall m \in \mathcal{P}(\mathbb{R}^d),\ \forall x \in \mathbb{R}^d, && |\nabla_x\phi(m,x)| + \bigl|\nabla_x^2\phi(m,x)\bigr| \leqslant C, \\ &\forall m \in \mathcal{P}(\mathbb{R}^d),\ \forall x, x' \in \mathbb{R}^d, && \biggl|\frac{\delta\phi}{\delta m}(m,x,x')\biggr| \leqslant C, \end{aligned}$$

for some $C \geqslant 0$. Under the settings of the theorem, let (X,m) be a random variable distributed as the stationary measure P. Then we have, for all $\phi \in \mathcal{C}_{\mathrm{b}}^{1,2}$,

$$\begin{aligned} \mathbb{E}\biggl[&\frac{1}{2}\Delta_x\phi(m,X) + b(m,X)\cdot\nabla_x\phi(m,X) \\ &+ \lambda\int \frac{\delta\phi}{\delta m}(m,X,x')(\delta_X - m)(\mathrm{d}x')\biggr] = 0. \end{aligned}$$

We omit the proof of the proposition, which follows directly from expanding the difference $\mathbb{E}[\phi(m_t,X_t)] - \mathbb{E}[\phi(m_0,X_0)]$ by Itō-type calculus.

The infinite-dimensional nature of the PDE above makes its analysis difficult, and in the following we approach the problem by studying a finite-dimensional projection of it. Under Assumption 7.11, define the functions

$$\beta(y,x) = -\nabla\Phi(y)\cdot\nabla\ell(x) = -\sum_{\nu=1}^{D}\nabla_\nu\Phi(y)\nabla\ell^\nu(x),$$
$$V(y,x) = \nabla\Phi(y)\cdot\ell(x) = \sum_{\nu=1}^{D}\nabla_\nu\Phi(y)\ell^\nu(x).$$

They verify $\beta(y,x) = -\nabla_x V(y,x)$. Note that, if m_y is a measure satisfying $\int \ell(x)m_y(\mathrm{d}x) = y$, then we have

$$\beta(y,x) = b(m_y,x),$$
$$V(y,x) = \frac{\delta F}{\delta m}(m_y,x).$$

Let (X,m) be distributed as the stationary measure P and consider the random variable $Y = \langle \ell, m\rangle = \int \ell(x)m(\mathrm{d}x)$ valued in $\mathcal{K}$. Applying Proposition 7.23 to functionals of the following form:

$$\phi(m,x) = \varphi\biggl(x, \int \ell(x')m(\mathrm{d}x')\biggr),$$

where $\varphi \in \mathcal{D}(\mathbb{R}^d\times\mathbb{R}^D)$, we get that the joint distribution $\rho = \rho^\lambda \coloneqq \mathrm{Law}(X,Y) \in \mathcal{P}(\mathbb{R}^d\times\mathcal{K})$ satisfies the stationary degenerate Fokker–Planck equation

$$\frac{1}{2}\Delta_x\rho - \nabla_x\cdot\bigl(\beta(y,x)\rho\bigr) - \lambda\nabla_y\cdot\Bigl(\bigl(\ell(x)-y\bigr)\rho\Bigr) = 0. \tag{7.13}$$

in the sense of distributions. The fact that P has finite fourth moment implies that its projection ρ satisfies the following moment condition:

$$\int (|x|^4 + |y|^2)\rho(\mathrm{d}x\,\mathrm{d}y) < \infty. \tag{7.14}$$

From the Fokker–Planck equation (7.13), we get the following result.

Lemma 7.24. *Under the setting of the theorem, for every function $\varphi \in \mathcal{C}^1(\mathbb{R}^D;\mathbb{R})$ whose gradient $\nabla\varphi$ is of linear growth, we have*

$$\iint \nabla\varphi(y)\cdot\bigl(\ell(x)-y\bigr)\rho(\mathrm{d}x\,\mathrm{d}y) = 0.$$

Proof of Lemma 7.24. Since its gradient $\nabla\varphi$ is of linear growth, the function φ is of quadratic growth. Thanks to the fact that ρ satisfies the moment condition (7.14), we can take the duality with φ in the static Fokker–Planck equation (7.13), from which the desired result follows. □

7.6.2 Entropy estimate

In this subsection, we show an entropy estimate for the stationary measure ρ by studying the Fokker–Planck equation (7.13).

Denote the first and second marginal distributions of ρ by ρ^1, ρ^2 respectively. For ρ^2-almost all $y \in \mathbb{R}^D$, denote also the conditional distribution of the first variable by $\rho^{1|2}(\cdot|y) : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}$. Define

$$\hat\rho(\mathrm{d}x\,\mathrm{d}y) \coloneqq \hat m_y(\mathrm{d}x)\rho^2(\mathrm{d}y) = \frac{1}{Z_y}\exp\bigl(-2V(y,x)\bigr)\,\mathrm{d}x\,\rho^2(\mathrm{d}y), \tag{7.15}$$

for $Z_y = \int \exp\bigl(-2V(y,x)\bigr)\,\mathrm{d}x$. Recall that $\hat m_y$ is the probability measure on $\mathbb{R}^d$ that satisfies a uniform LSI according to Assumption 7.12.

Proposition 7.25. *Under the setting of the theorem, we have*

$$H(\rho|\hat\rho) \leqslant \frac{C_{\mathrm{LS}}}{2}\biggl(D + 2\iint \bigl(\ell(x)-y\bigr)^\top \nabla^2\Phi(y)\bigl(\ell(x)-y\bigr)\rho(\mathrm{d}x\,\mathrm{d}y)\biggr)\lambda,$$

where $\hat\rho$ is the measure defined by (7.15).

For convenience, we set

$$I \coloneqq \iint \bigl(\ell(x)-y\bigr)^\top \nabla^2\Phi(y)\bigl(\ell(x)-y\bigr)\rho(\mathrm{d}x\,\mathrm{d}y), \tag{7.16}$$

so the claim of the proposition reads

$$H(\rho|\hat\rho) \leqslant \frac{C_{\mathrm{LS}}}{2}(D+2I)\lambda.$$

Proof of Proposition 7.25. Let g^ε be the Gaussian kernel in D dimensions:

$$g^\varepsilon(y) = (2\pi\varepsilon)^{-D/2}\exp\biggl(-\frac{|y|^2}{2\varepsilon}\biggr).$$

We define the partial convolution $\rho^\varepsilon = \rho \star_y g^\varepsilon$, and according to (7.13), it solves the non-degenerate elliptic equation

$$\frac12\Delta_x\rho^\varepsilon - \nabla_x\cdot\bigl(\beta^\varepsilon(y,x)\rho^\varepsilon\bigr) - \ell(x)\cdot\nabla_y\rho^\varepsilon + \lambda\varepsilon\Delta_y\rho^\varepsilon + \lambda\nabla_y\cdot(y\rho^\varepsilon) = 0, \tag{7.17}$$

where β^ε is defined by

$$\beta^\varepsilon = \frac{(\beta\rho)\star_y g^\varepsilon}{\rho^\varepsilon}.$$

Indeed, we have

$$\begin{aligned}
(y\rho)\star_y g^\varepsilon &= \int y'\rho(x',y')g^\varepsilon(y-y')\,\mathrm{d}y' \\
&= \int (y'-y+y)\rho(x',y')g^\varepsilon(y-y')\,\mathrm{d}y' \\
&= \varepsilon\int \rho(x',y')\nabla_y g^\varepsilon(y-y')\,\mathrm{d}y' + y\rho^\varepsilon \\
&= \varepsilon\nabla_y\rho^\varepsilon + y\rho^\varepsilon.
\end{aligned}$$

Thus, $\bigl(\nabla_y\cdot(y\rho)\bigr)\star_y g^\varepsilon = \nabla_y\cdot\bigl((y\rho)\star_y g^\varepsilon\bigr) = \varepsilon\Delta_y\rho^\varepsilon + \nabla_y\cdot(y\rho^\varepsilon)$. By [22, Lemma 3.1.1], we have $\|\beta^\varepsilon\|_{L^2(\rho^\varepsilon)} \leqslant \|\beta\|_{L^2(\rho)} < \infty$. Then we can apply [22, Theorem 3.1.2] to (7.17) and obtain that $\rho^\varepsilon \in W^{1,1}(\mathbb{R}^{d+D})$ and satisfies

$$\frac{1}{2}\iint \frac{|\nabla_x \rho^\varepsilon|^2}{\rho^\varepsilon} + \lambda\varepsilon \iint \frac{|\nabla_y \rho^\varepsilon|^2}{\rho^\varepsilon} = \iint \nabla_x \rho^\varepsilon \cdot \beta^\varepsilon + \lambda \iint \nabla_y \rho^\varepsilon \cdot \ell - \lambda \iint \nabla_y \rho^\varepsilon \cdot y.$$

As the function ℓ depends only on the x argument, for the second term we have

$$\iint \nabla_y \rho^\varepsilon \cdot \ell = \int \Big(\int \nabla_y \rho^\varepsilon(x,y) dy \Big) \ell(x)\, \mathrm{d}x = \int 0 \cdot \ell(x)\, \mathrm{d}x = 0.$$

where the first equality is due to Fubini and the second to the fact that $\nabla \rho^\varepsilon \in L^1_x(L^1_y)$. For the last term, similarly, since the function $\big((x,y) \mapsto \rho^\varepsilon(x,y)y\big) \in W^{1,1}$, we have $\iint \nabla_y \cdot (\rho^\varepsilon y) = 0$ and therefore,

$$-\iint \nabla_y \rho^\varepsilon \cdot y = \iint (\nabla_y \cdot y)\rho^\varepsilon = D.$$

That is to say, we have established

$$\frac{1}{2}\iint \frac{|\nabla_x \rho^\varepsilon|^2}{\rho^\varepsilon} + \lambda\varepsilon \iint \frac{|\nabla_y \rho^\varepsilon|^2}{\rho^\varepsilon} = \iint \nabla_x \rho^\varepsilon \cdot \beta^\varepsilon + \lambda D. \tag{7.18}$$

This equality implies a uniform-in-ε bound on $\iint |\nabla_x \rho^\varepsilon|^2/\rho^\varepsilon$ by Cauchy–Schwarz. Using a sequence of functions in $\mathcal{C}^\infty_\mathrm{c}$ approaching $V(y,x)$ in (7.13), we also get

$$\iint \frac{1}{2}\beta \cdot \nabla_x \rho^\varepsilon - \iint \lambda\varepsilon \nabla_y V \cdot \nabla_y \rho^\varepsilon - \iint \beta \cdot \beta^\varepsilon \rho^\varepsilon + \iint \lambda \nabla_y V(y,x) \cdot \big(\ell(x) - y\big)\rho^\varepsilon(\mathrm{d}x\,\mathrm{d}y) = 0. \tag{7.19}$$

The second term of (7.19) is upper bounded by

$$\begin{aligned}\lambda\varepsilon \iint |\nabla_y V \cdot \nabla_y \rho^\varepsilon| &\leqslant \lambda\varepsilon \|\nabla_y V\|_{L^2(\rho^\varepsilon)} \Big(\iint \frac{|\nabla_y \rho^\varepsilon|^2}{\rho^\varepsilon} \Big)^{1/2} \\ &\leqslant \sqrt{\lambda\varepsilon}\|\nabla_y V\|_{L^2(\rho^\varepsilon)} \sqrt{\frac{1}{2}\iint \frac{|\nabla_x \rho^\varepsilon|^2}{\rho^\varepsilon} + \frac{1}{2}\|\beta^\varepsilon\|^2_{L^2(\rho^\varepsilon)} + \lambda D},\end{aligned}$$

where the second inequality is due to (7.18). Using the uniform-in-ε bound of $\iint |\nabla_x \rho^\varepsilon|^2/\rho^\varepsilon$, we obtain that the second term of (7.19) converges to 0 when $\varepsilon \to 0$. The third term of (7.19) satisfies

$$\iint \beta \cdot \beta^\varepsilon \rho^\varepsilon = \iint \beta \cdot \big((\beta\rho) \star g^\varepsilon\big) = \iint (\beta \star g^\varepsilon) \cdot \beta\rho \to \iint |\beta|^2 \rho$$

when $\varepsilon \to 0$. Hence, by (7.18) and (7.19), we have

$$\begin{aligned}&\frac{1}{2}\iint |\nabla_x \log \rho^\varepsilon - 2\beta|^2 \rho^\varepsilon(\mathrm{d}x\,\mathrm{d}y) \\ &\quad\leqslant \lambda D + \iint \nabla_x \rho^\varepsilon \cdot \beta^\varepsilon + \iint \beta \cdot \nabla_x \rho^\varepsilon - 2\iint \nabla_x \rho^\varepsilon \cdot \beta \\ &\qquad + 2\lambda \iint \nabla_y V(y,x) \cdot \big(\ell(x) - y\big)\rho^\varepsilon(\mathrm{d}x\,\mathrm{d}y) + o(1) \\ &\quad= \lambda D + 2\lambda \iint \nabla_y V(y,x) \cdot \big(\ell(x) - y\big)\rho^\varepsilon(\mathrm{d}x\,\mathrm{d}y) + o(1),\end{aligned}$$

where the last equality is due to the fact that

$$\left|\iint \nabla_x \rho^\varepsilon \cdot (\beta^\varepsilon - \beta)\right| \leqslant \left(\iint \frac{|\nabla_x \rho^\varepsilon|^2}{\rho^\varepsilon}\right)^{1/2} \left(\iint |\beta - \beta^\varepsilon|^2 \rho^\varepsilon\right)^{1/2}$$

and

$$\begin{aligned}\iint |\beta - \beta^\varepsilon|^2 \rho^\varepsilon &= \iint |\beta|^2 \rho^\varepsilon - 2 \iint \beta \cdot \beta^\varepsilon \rho^\varepsilon + \iint |\beta^\varepsilon|^2 \rho^\varepsilon \\ &= \iint (|\beta|^2 \star_y g^\varepsilon)\rho - 2 \iint (\beta \star_y g^\varepsilon) \cdot \beta\rho + \iint |\beta^\varepsilon|^2 \rho^\varepsilon \to 0,\end{aligned}$$

when $\varepsilon \to 0$ by previous arguments and [22, Lemma 3.1.1]. We also have

$$\begin{aligned}\iint \nabla_y V(y,x) \cdot \big(\ell(x) - y\big)\rho^\varepsilon(\mathrm{d}x\,\mathrm{d}y) &= \iint \Big(\nabla_y V(y,x) \cdot \big(\ell(x) - y\big)\Big) \star_y g^\varepsilon \rho(\mathrm{d}x\,\mathrm{d}y) \\ &\to \iint \nabla_y V(y,x) \cdot \big(\ell(x) - y\big)\rho(\mathrm{d}x\,\mathrm{d}y)\end{aligned}$$

when $\varepsilon \to 0$. Finally, by Lemma 7.24 for the function $\varphi(y) = y \cdot \nabla\Phi(y) - \Phi(y)$, we have

$$\begin{aligned}&\iint \nabla_y V(y,x) \cdot \big(\ell(x) - y\big)\rho(\mathrm{d}x\,\mathrm{d}y) \\ &\quad= \iint \ell(x)^\top \nabla^2\Phi(y)\big(\ell(x) - y\big)\rho(\mathrm{d}x\,\mathrm{d}y) \\ &\quad= \iint \big(\ell(x) - y\big)^\top \nabla^2\Phi(y)\big(\ell(x) - y\big)\rho(\mathrm{d}x\,\mathrm{d}y) = I,\end{aligned}$$

where the last equality is exactly the definition (7.16) of I. Thus, we have shown

$$\frac{1}{2}\iint \left|\nabla_x \log \frac{\rho^\varepsilon(x,y)}{\hat m_y(x)}\right|^2 \rho^\varepsilon(\mathrm{d}x\,\mathrm{d}y) \leqslant (D + 2I)\lambda + o(1).$$

Note that by the lower semicontinuity of (partial) Fisher information,

$$\liminf_{\varepsilon \to 0} \iint \left|\nabla_x \log \frac{\rho^\varepsilon(x,y)}{\hat m_y(x)}\right|^2 \rho^\varepsilon(\mathrm{d}x\,\mathrm{d}y) \geqslant \iint \left|\nabla_x \log \frac{\rho(x,y)}{\hat m_y(x)}\right|^2 \rho(\mathrm{d}x\,\mathrm{d}y).$$

We refer readers to the proof of Lemma B.1 for details. Taking the limit $\varepsilon \to 0$, we obtain

$$\frac{1}{2}\iint \left|\nabla_x \log \frac{\rho(x,y)}{\hat m_y(x)}\right|^2 \rho(\mathrm{d}x\,\mathrm{d}y) \leqslant (D + 2I)\lambda.$$

Since ρ^2 is supported on $\mathcal{K}$, for ρ^2-almost all $y \in \mathbb{R}^D$, we have the following by the uniform LSI for $(\hat m_y)_{y\in\mathcal{K}}$:

$$H(\rho|\hat\rho) = \int H\big(\rho^{1|2}(\cdot|y)\big|\hat m_y\big)\rho^2(\mathrm{d}y) \leqslant \frac{C_{\mathrm{LS}}}{4}\iint \left|\nabla_x \log \frac{\rho(x,y)}{\hat m_y(x)}\right|^2 \rho(\mathrm{d}x\,\mathrm{d}y),$$

which completes the proof. □

Remark 7.26. If we formally integrate the static Fokker–Planck equation (7.13) with $\log(\rho/\hat\rho)$ and integrate by parts, we obtain

$$\frac{1}{2}\iint\left|\nabla_x \log\frac{\rho(x,y)}{\hat m_y(x)}\right|^2\rho(\mathrm{d}x\,\mathrm{d}y) = \lambda D + 2\lambda\iint \ell(x)^\top\nabla^2\Phi(y)\bigl(\ell(x)-y\bigr)\rho(\mathrm{d}x\,\mathrm{d}y). \tag{7.20}$$

However, the equality must not hold in all circumstances. Indeed, if one artificially increases the dimension of ℓ and Φ by defining the new functions

$$\begin{aligned}\tilde\Phi(y_0,y_1) &= \Phi(y_0),\\ \tilde\ell(x) &= \bigl(\ell(x),0\bigr),\end{aligned}$$

the right hand side of (7.20) increases while the left hand side stays unchanged. This phenomenon is caused by the fact that the equation (7.13) is degenerate elliptic and lacks Laplacian in the y directions. To illustrate this effect, consider the first-order equation

$$\nabla_y\cdot(y\rho) = 0$$

in d dimensions. This equation has a probability solution $\rho=\delta_0$, the Dirac mass at the origin. Formally integrating the equation with $\log\rho$ and integrating by parts, we have

$$0 = \int\log\rho\nabla_y\cdot(y\rho) = -\int\frac{\nabla_y\rho}{\rho}\cdot y\rho = -\int\nabla_y\rho\cdot y = \int\rho\nabla_y\cdot y = \int\rho d = d,$$

which is absurd.

To complete the entropy estimate, we provide in the following upper bounds for the integral I.

Proposition 7.27. *Under the setting of the theorem, the integral I in (7.16) satisfies the upper bound:*

$$I \leqslant 12M_2\Bigl(dC_{\mathrm{LS}} + 2W_2^2\bigl(\mathrm{Law}(X),m_*\bigr)\Bigr), \tag{7.21}$$

where X is the first component of the random variable (X,m) following the stationary distribution $P=P^\lambda$, and m_* is the unique invariant measure of the McKean–Vlasov process (7.1). If additionally $\nabla^2\Phi$ is convex and the quantity

$$M_1 \coloneqq \sup_{x\in\mathbb{R}^d,y\in\mathbb{R}^D}\ell(x)^\top\nabla^2\Phi(y)\ell(x)$$

is finite, then we have the alternative upper bound:

$$I\leqslant M_1. \tag{7.22}$$

Proof of Proposition 7.27. First let us treat the simpler case where $M_1<\infty$. By applying Lemma 7.24 to the function $\varphi(y) = y\cdot\nabla\Phi(y)-\Phi(y)$, we get

$$\begin{aligned}I &= \iint\bigl(\ell(x)^\top\nabla^2\Phi(y)\ell(x) - y^\top\nabla^2\Phi(y)y\bigr)\rho(\mathrm{d}x\,\mathrm{d}y)\\ &\leqslant \iint\ell(x)^\top\nabla^2\Phi(y)\ell(x)\rho(\mathrm{d}x\,\mathrm{d}y)\leqslant M_1.\end{aligned}$$

So the second claim of the proposition is proved.

Without the assumption $M_1 < \infty$, we note that, for the second-order functional derivative

$$\frac{\delta^2 F}{\delta m^2}(m, x', x'') = \ell(x'')^\top \nabla^2 \Phi(\langle \ell, m\rangle)\ell(x),$$

we have

$$I = \mathbb{E}\Big[\iint \frac{\delta^2 F}{\delta m^2}(m, x', x'')(\delta_X - m)(\mathrm{d}x')(\delta_X - m)(\mathrm{d}x'')\Big],$$

where (X, m) is a random variable following the stationary distribution P, and $(X, \langle \ell, m\rangle)$ has the distribution ρ. We observe

$$\begin{aligned}\big|D_m^2 F(m, x', x'')\big| &= \Big|\nabla_{x'}\nabla_{x''}\frac{\delta^2 F}{\delta m^2}(m, x', x'')\Big| \\ &= \big|\nabla\ell(x'')^\top \nabla^2\Phi(\langle \ell, m\rangle)\nabla\ell(x')\big| \leqslant M_2.\end{aligned}$$

Then, by applying Lemma D.1 in Appendix D.1 to a sequence of $\mathcal{C}^2$ functions approaching $(x', x'') \mapsto \frac{\delta^2 F}{\delta m^2}(m, x', x'')$, we get

$$\begin{aligned}I &\leqslant M_2\,\mathbb{E}\big[W_2^2(\delta_X, m)\big] \\ &= M_2\,\mathbb{E}\Big[\int |X - x'|^2 m(\mathrm{d}x')\Big] \\ &\leqslant 2M_2\,\mathbb{E}\Big[|X - \mathbb{E}[X]|^2 + \int |x' - \mathbb{E}[X]|^2 m(\mathrm{d}x')\Big].\end{aligned}$$

Let $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be the functional defined by

$$\phi(m) = \int |x' - \mathbb{E}[X]|^2 m(\mathrm{d}x').$$

We consider the sequence of $\mathcal{C}^1_{\mathrm{b}}$ "soft cut-offs" that approach ϕ:

$$\phi_n(m) = \sum_{i=1}^d \int n^2 \tanh^2\Big(\frac{x'^i - \mathbb{E}[X^i]}{n}\Big) m(\mathrm{d}x'), \qquad \text{for } n \in \mathbb{N}.$$

Then, by applying the sequence ϕ_n to Proposition 7.23 of stationary measure and taking the limit $n \to \infty$, we get

$$\begin{aligned}0 = \lambda\,\mathbb{E}\Big[\int \frac{\delta\phi}{\delta m}(m, x')(\delta_X - m)(\mathrm{d}x')\Big] \\ = \lambda\,\mathbb{E}\big[|X - \mathbb{E}[X]|^2\big] - \lambda\,\mathbb{E}\Big[\int |x' - \mathbb{E}[X]|^2 m(\mathrm{d}x')\Big].\end{aligned}$$

Thus, we have derived

$$I \leqslant 4M_2\,\mathbb{E}\big[|X - \mathbb{E}[X]|^2\big] =: 4M_2 \operatorname{Var} X,$$

where $\operatorname{Var} X$ denotes the sum of the variances of each component of the random vector X. It remains only to find an upper bound for $\operatorname{Var} X$. Note that, using the definition of Wasserstein distance and the triangle inequality, and letting X_* be distributed as m_*, we get

$$\begin{aligned}\operatorname{Var} X &= W_2^2\big(\mathrm{Law}(X), \delta_{\mathbb{E}[X]}\big)\\ &\leqslant 3\Big(W_2^2\big(\mathrm{Law}(X), m_*\big) + W_2^2\big(m_*, \delta_{\mathbb{E}[X_*]}\big) + W_2^2\big(\delta_{\mathbb{E}[X_*]}, \delta_{\mathbb{E}[X]}\big)\Big)\\ &\leqslant 3\Big(\operatorname{Var} X_* + 2W_2^2\big(\mathrm{Law}(X), m_*\big)\Big),\end{aligned}$$

while the variance of X_* is upper bounded by the Poincaré inequality:

$$\begin{aligned}\operatorname{Var} X_* &= \sum_{i=1}^d \bigg(\int |x^i|^2 m_*(\mathrm{d}x) - \Big(\int x^i m_*(\mathrm{d}x)\Big)^2\bigg)\\ &\leqslant C_{\mathrm{LS}} \sum_{i=1}^d \int |\nabla x^i|^2 m_*(\mathrm{d}x) = C_{\mathrm{LS}} d.\end{aligned}$$

We then conclude by combining the three inequalities above. □

7.6.3 Construction of another measure

In this subsection, we construct another measure in order to exploit the convexity of Φ, used for the proof of the first and second claims of the theorem. Readers only interested in the last claim of the theorem can now directly go to the next subsection.

Let $\mu = \mu^\lambda$ be the probability measure on $\mathbb{R}^d \times \mathbb{R}^D$ characterized by the following formula:

$$\begin{aligned}\langle f, \mu\rangle &= \int f(x,y)\mu(\mathrm{d}x\,\mathrm{d}y)\\ &= \mathbb{E}\bigg[\int f(x, \langle \ell, m\rangle) m(\mathrm{d}x)\bigg] = \mathbb{E}\bigg[\int f(x, Y) m(\mathrm{d}x)\bigg] \qquad (7.23)\end{aligned}$$

for all bounded and measurable $f : \mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$. By taking f depending only on the y variable, we first realize that the second marginals of ρ and μ agree, that is,

$$\rho^2 = \mu^2.$$

In addition, we show the following important properties of μ.

Proposition 7.28. *Under the setting of the theorem, for every $\mathcal{C}^2$ differentiable $\Psi : \mathbb{R}^D \to \mathbb{R}$ with bounded Hessian, we have*

$$\iint \nabla\Psi(y) \cdot \ell(x)(\mu - \rho)(\mathrm{d}x\,\mathrm{d}y) = 0.$$

In particular, the respective first marginals μ^1, ρ^1 of μ, ρ satisfy

$$\int \ell(x)(\mu^1 - \rho^1)(\mathrm{d}x) = 0.$$

Moreover, denoting by $\mu^{1|2}(\cdot|\cdot) : \mathcal{B}(\mathbb{R}^d) \times \mathbb{R}^D \to \mathbb{R}$ the conditional measure of μ given its second variable, we have for ρ^2-almost all $y \in \mathbb{R}^D$,

$$\big\langle \ell, \mu^{1|2}(\cdot|y)\big\rangle = y.$$

Proof of Proposition 7.28. Consider the functional

$$\phi(m) = \int f(x', \langle k, m\rangle) m(\mathrm{d}x'),$$

where $f \in \mathcal{C}^1_\mathrm{b}(\mathbb{R}^d \times \mathbb{R}^D; \mathbb{R})$ and $k \in \mathcal{C}_\mathrm{b}(\mathbb{R}^d; \mathbb{R}^D)$. Then its linear functional derivative reads

$$\frac{\delta\phi}{\delta m}(m, x') = f(x', \langle k, m\rangle) + \int \nabla_y f(x'', \langle k, m\rangle) \cdot k(x') m(\mathrm{d}x''),$$

so ϕ belongs to the $\mathcal{C}^1_\mathrm{b}$ class. Then, applying Proposition 7.23 to the functional ϕ, we get

$$\begin{aligned} 0 &= \mathbb{E}\biggl[\int \frac{\delta\phi}{\delta m}(m, x')(\delta_X - m)(\mathrm{d}x')\biggr] \\ &= \mathbb{E}[f(X, m)] - \mathbb{E}\biggl[\int f(x', \langle k, m\rangle) m(\mathrm{d}x')\biggr] \\ &\quad + \mathbb{E}\biggl[\int \nabla_y f(x', \langle k, m\rangle) \cdot \bigl(k(X) - \langle k, m\rangle\bigr) m(\mathrm{d}x')\biggr] \end{aligned}$$

By approximation, the equality above holds for $k = \ell$ and for all $\mathcal{C}^1$-continuous f with the following growth bounds:

$$\begin{aligned} |f(x, y)| &\leqslant M(1 + |x|^4 + |y|^2), \\ |\nabla_y f(x, y)| &\leqslant M(1 + |x|^2 + |y|), \end{aligned}$$

that is to say, we have

$$\langle f, \rho - \mu\rangle + \mathbb{E}\biggl[\int \nabla_y f(x', Y)\bigl(\ell(X) - Y\bigr) m(\mathrm{d}x')\biggr] = 0,$$

where, as before, $Y = \langle \ell, m\rangle$. Specializing to $f(x, y) = \nabla\Psi(y) \cdot \ell(x)$, we obtain

$$\begin{aligned} \langle f, \mu - \rho\rangle &= \mathbb{E}\biggl[\int \ell(x)^\top \nabla^2\Psi(Y)\bigl(\ell(X) - Y\bigr) m(\mathrm{d}x)\biggr] \\ &= \mathbb{E}\bigl[Y^\top \nabla^2\Psi(Y)\bigl(\ell(X) - Y\bigr)\bigr] = 0, \end{aligned}$$

where the last equality is due to Lemma 7.24, as for $\varphi(y) \coloneqq \nabla\Psi(y) \cdot y - \Psi(y)$, we have $\nabla\varphi(y) = \nabla^2\Psi(y)y$. So the first claim is proved. Taking $\Psi(y) = y^\nu$, for $\nu = 1, \ldots, D$, yields the second claim.

For the last claim, we take $f(x, y) = \ell(x)g(y)$ for $g : \mathbb{R}^D \to \mathbb{R}$ of linear growth in the defining equation (7.23) of μ. Then, we get

$$\begin{aligned} \int g(y)\biggl(\int \ell(x)\mu^{1|2}(\mathrm{d}x|y)\biggr)\mu^2(\mathrm{d}y) &= \iint f(x, y)\mu(\mathrm{d}x\,\mathrm{d}y) = \mathbb{E}[Yg(Y)] \\ &= \int g(y) y \mu^2(\mathrm{d}y). \end{aligned}$$

The desired property follows from the arbitrariness of g. $\square$

7.6.4 Proving the theorem

Having established the entropy estimate and constructed the auxiliary measure, we finally move to the central part of the proof, which consists of six steps. The aim of the first five steps is to show the first and the second claims of the theorem, and in the last step we prove the last claim.

Step 1: Control of the symmetrized entropy. We aim at controlling the symmetrized entropy

$$\int \big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y)$$

in this step. First observe

$$\begin{aligned}\int \big(H(\hat{m}_y|m_*) &+ H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y)\\ &= 2\iint \big(V(y_*,x) - V(y,x)\big)\big(\hat{m}_y(\mathrm{d}x) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y). \end{aligned} \tag{7.24}$$

In order to control the right hand side above, we turn to the probability measure μ introduced in (7.23). Recall that m_* is the invariant measure of the McKean–Vlasov (7.1), and $y_* := \langle \ell, m_*\rangle$. The convexity of Φ implies the convexity of F as a functional, and as a result, for ρ^2-almost all $y \in \mathbb{R}^D$, we have the tangent inequalities

$$\begin{aligned}\int V\big(\langle \ell, \mu^{1|2}(\cdot|y)\rangle, x\big)&\big(\mu^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\\ &\geqslant F\big(\mu^{1|2}(\cdot|y)\big) - F(m_*)\\ &\geqslant \int V(y_*,x)\big(\mu^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big). \end{aligned} \tag{7.25}$$

Thanks to the last claim of Proposition 7.28, the leftmost term satisfies, for ρ^2-almost all $y \in \mathbb{R}^D$,

$$\int V\big(\langle \ell, \mu^{1|2}(\cdot|y)\rangle, x\big)\big(\mu^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big) = \int V(y,x)\big(\mu^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big).$$

Hence, integrating the tangent inequalities (7.25) above by ρ^2, we get

$$\begin{aligned}\iint V(y,x)&\big(\mu^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y)\\ &\geqslant \iint V(y_*,x)\big(\mu^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y). \end{aligned} \tag{7.26}$$

Using $\mu^2 = \rho^2$ and applying Proposition 7.28 to $V(y,x) = \nabla\Phi(y)\cdot\ell(x)$, we know that the left hand side of (7.26) satisfies

$$\begin{aligned}&\iint V(y,x)\big(\mu^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y)\\ &\quad= \iint V(y,x)\mu(\mathrm{d}x\,\mathrm{d}y) - \iint V(y,x)m_*(\mathrm{d}x)\rho^2(\mathrm{d}y)\\ &\quad= \iint V(y,x)\rho(\mathrm{d}x\,\mathrm{d}y) - \iint V(y,x)m_*(\mathrm{d}x)\rho^2(\mathrm{d}y)\\ &\quad= \iint V(y,x)\big(\rho^{1|2}(\mathrm{d}x|y) - m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y). \end{aligned}$$

The right hand side of (7.26) satisfies

$$\begin{aligned}
&\iint V(y_*,x)\big(\mu^{1|2}(\mathrm{d}x|y)-m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y)\\
&\quad=\iint V(y_*,x)\mu(\mathrm{d}x\,\mathrm{d}y)-\iint V(y_*,x)m_*(\mathrm{d}x)\\
&\quad=\nabla\Phi(y_*)\cdot\int\ell(x)\big(\mu^1(\mathrm{d}x)-m_*(\mathrm{d}x)\big)\\
&\quad=\nabla\Phi(y_*)\cdot\int\ell(x)\big(\rho^1(\mathrm{d}x)-m_*(\mathrm{d}x)\big)\\
&\quad=\iint V(y_*,x)\big(\rho^{1|2}(\mathrm{d}x|y)-m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y),
\end{aligned}$$

where the third equality is due to the last claim of Proposition 7.28. Thus, we have derived

$$\iint\big(V(y,x)-V(y_*,x)\big)\big(\rho^{1|2}(\mathrm{d}x|y)-m_*(\mathrm{d}x)\big)\rho^2(\mathrm{d}y)\geqslant 0. \tag{7.27}$$

Therefore, to dominate the right hand side of (7.24), it remains to control the following term. Using the Kantorovich duality, we get

$$\begin{aligned}
&\left|\iint\big(V(y,x)-V(y_*,x)\big)\big(\rho^{1|2}(\mathrm{d}x|y)-\hat m_y(\mathrm{d}x)\big)\rho^2(\mathrm{d}y)\right|\\
&\quad=\left|\int\big(\nabla\Phi(y)-\nabla\Phi(y_*)\big)\cdot\big\langle\ell,\rho^{1|2}(\cdot|y)-\hat m_y\big\rangle\rho^2(\mathrm{d}y)\right|\\
&\quad=\left|\iint_0^1(y-y_*)^\top\nabla^2\Phi\big((1-t)y+ty_*\big)\big\langle\ell,\rho^{1|2}(\cdot|y)-\hat m_y\big\rangle\,\mathrm{d}t\,\rho^2(\mathrm{d}y)\right|\\
&\quad\leqslant\left(\iint_0^1(y-y_*)^\top\nabla^2\Phi\big((1-t)y+ty_*\big)(y-y_*)\,\mathrm{d}t\,\rho^2(\mathrm{d}y)\right)^{1/2}\\
&\qquad\times\left(\iint_0^1\left|\Big\langle\nabla^2\Phi\big((1-t)y+ty_*\big)^{1/2}\ell,\rho^{1|2}(\cdot|y)-\hat m_y\Big\rangle\right|^2\mathrm{d}t\,\rho^2(\mathrm{d}y)\right)^{1/2}\\
&\quad\leqslant\left(\iint_0^1(y-y_*)^\top\nabla^2\Phi\big((1-t)y+ty_*\big)(y-y_*)\,\mathrm{d}t\,\rho^2(\mathrm{d}y)\right)^{1/2}\\
&\qquad\times\sqrt{M_2}\left(\int W_1^2\big(\rho^{1|2}(\cdot|y),\hat m_y\big)\rho^2(\mathrm{d}y)\right)^{1/2}\\
&\quad=:\sqrt{M_2 v(\rho^2)}\left(\int W_1^2\big(\rho^{1|2}(\cdot|y),\hat m_y\big)\rho^2(\mathrm{d}y)\right)^{1/2},
\end{aligned}$$

where, by Assumption 7.14, $\big|\nabla^2\Phi(y')^{1/2}\nabla\ell(x)\big|\leqslant\sqrt{M_2}$ for all $x\in\mathbb{R}^d$, $y'\in\mathbb{R}^D$, and $v(\rho^2)$ is the quantity to be controlled in the first claim of the theorem. The uniform LSI for $(\hat m_y)_{y\in\mathcal{K}}$ implies a uniform Talagrand's transport inequality, from which we obtain

$$\begin{aligned}
\int W_1^2\big(\rho^{1|2}(\cdot|y),\hat m_y\big)\rho^2(\mathrm{d}y)&\leqslant\int W_2^2\big(\rho^{1|2}(\cdot|y),\hat m_y\big)\rho^2(\mathrm{d}y)\\
&\leqslant C_{\mathrm{LS}}\int H\big(\rho^{1|2}(\cdot|y)\big|\hat m_y\big)\rho^2(\mathrm{d}y)=C_{\mathrm{LS}}H(\rho|\hat\rho),
\end{aligned}$$

as ρ^2 is supported on $\mathcal{K}$. Combining the two inequalities above with (7.24) and (7.27), we obtain

$$\int \big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y) \leqslant 2\sqrt{M_2 C_{\mathrm{LS}} v(\rho^2) H(\rho|\hat{\rho})}. \tag{7.28}$$

Step 2: Control of the conditional Wasserstein distance. Now, using again the Talagrand's transport inequality for $\hat{m}_y$ and m_* (note that $m_* = \hat{m}_{y_*}$ for $y_* = \langle \ell, m_* \rangle$), we get for ρ^2-almost all $y \in \mathbb{R}^D$,

$$W_2^2(\hat{m}_y, m_*) \leqslant \frac{C_{\mathrm{LS}}}{2}\big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big),$$

while the triangle inequality and the transport inequality imply

$$\begin{aligned} W_2^2\big(\rho^{1|2}(\cdot|y), m_*\big) &\leqslant 2\Big(W_2^2\big(\rho^{1|2}(\cdot|y), \hat{m}_y\big) + W_2^2(\hat{m}_y, m_*)\Big) \\ &\leqslant 2C_{\mathrm{LS}} H\big(\rho^{1|2}(\cdot|y), \hat{m}_y\big) + 2W_2^2(\hat{m}_y, m_*). \end{aligned}$$

So, combining the three inequalities above and integrating with ρ^2, we find

$$\int W_2^2\big(\rho^{1|2}(\cdot|y), m_*\big)\rho^2(dy) \leqslant 2C_{\mathrm{LS}} H(\rho|\hat{\rho}) + 2\sqrt{M_2 C_{\mathrm{LS}}^3 v(\rho^2) H(\rho|\hat{\rho})}. \tag{7.29}$$

Step 3: Control of $v(\rho^2)$ by $H(\rho|\hat{\rho})$. Applying Proposition 7.28 to the function $\Psi(y) = \Phi(y) - \nabla\Phi(y_*) \cdot y$, where, as we recall, $y_* = \langle \ell, m_* \rangle$, we get

$$\begin{aligned} 0 &= \iint \nabla\Psi(y) \cdot \ell(x)(\mu - \rho)(\mathrm{d}x\,\mathrm{d}y) \\ &= \iint \big(\nabla\Phi(y) - \nabla\Phi(y_*)\big) \cdot \ell(x)(\mu - \rho)(\mathrm{d}x\,\mathrm{d}y) \\ &= \iint_0^1 (y - y_*)^\top \nabla^2\Phi\big((1-t)y + ty_*\big)\big\langle \ell, \mu^{1|2}(\cdot|y) - \rho^{1|2}(\cdot|y)\big\rangle\,\mathrm{d}t\,\rho^2(\mathrm{d}y) \\ &= \iint_0^1 (y - y_*)^\top \nabla^2\Phi\big((1-t)y + ty_*\big)\big(y - \big\langle \ell, \rho^{1|2}(\cdot|y)\big\rangle\big)\,\mathrm{d}t\,\rho^2(\mathrm{d}y), \end{aligned}$$

where the last equality is due to the last claim of Proposition 7.28. In other words, we have

$$\begin{aligned} &\iint_0^1 (y - y_*)^\top \nabla^2\Phi\big((1-t)y + ty_*\big)(y - y_*)\,\mathrm{d}t\,\rho^2(\mathrm{d}y) \\ &\qquad = \iint_0^1 (y - y_*)^\top \nabla^2\Phi\big((1-t)y + ty_*\big)\big\langle \ell, \rho^{1|2}(\cdot|y) - m_*\big\rangle\,\mathrm{d}t\,\rho^2(\mathrm{d}y), \end{aligned}$$

and this implies, by Cauchy–Schwarz,

$$\begin{aligned} &\iint_0^1 (y - y_*)^\top \nabla^2\Phi\big((1-t)y + ty_*\big)(y - y_*)\,\mathrm{d}t\,\rho^2(\mathrm{d}y) \\ &\quad\leqslant \iint_0^1 \big\langle \ell, \rho^{1|2}(\cdot|y) - m_*\big\rangle^\top \nabla^2\Phi\big((1-t)y + ty_*\big)\big\langle \ell, \rho^{1|2}(\cdot|y) - m_*\big\rangle\,\mathrm{d}t\,\rho^2(\mathrm{d}y) \\ &\quad= \iint_0^1 \Big|\Big\langle \nabla^2\Phi\big((1-t)y + ty_*\big)^{1/2}\ell, \rho^{1|2}(\cdot|y) - m_*\Big\rangle\Big|^2\,\mathrm{d}t\,\rho^2(\mathrm{d}y). \end{aligned}$$

As $\left|\nabla^2\Phi(y')^{1/2}\nabla\ell(x)\right| \leqslant \sqrt{M_2}$ for all $y' \in \mathbb{R}^D$ and $x \in \mathbb{R}^d$, we have, by the Kantorovich duality,

$$
\begin{aligned}
v(\rho^2) &= \iint_0^1 (y-y_*)^\top \nabla^2\Phi\big((1-t)y+ty_*\big)(y-y_*)\,\mathrm{d}t\,\rho^2(\mathrm{d}y) \\
&\leqslant M_2 \int W_1^2\big(\rho^{1|2}(\cdot|y), m_*\big)\rho^2(\mathrm{d}y). \qquad (7.30)
\end{aligned}
$$

Thus, using the fact that $W_1 \leqslant W_2$ and the inequality (7.29), we obtain

$$
v(\rho^2) \leqslant 2M_2C_{\mathrm{LS}}H(\rho|\hat{\rho}) + 2\sqrt{M_2^3C_{\mathrm{LS}}^3 v(\rho^2)H(\rho|\hat{\rho})}.
$$

Introduce the "adimensionalized" variable

$$
\upsilon = \frac{v(\rho^2)}{4M_2^3C_{\mathrm{LS}}^3H(\rho|\hat{\rho})}.
$$

Then the inequality above reads

$$
\upsilon \leqslant \frac{1}{2M_2^2C_{\mathrm{LS}}^2} + \sqrt{\upsilon} \leqslant \frac{1}{2M_2^2C_{\mathrm{LS}}^2} + \frac{1}{2} + \frac{1}{2}\upsilon.
$$

Hence, we get

$$
v(\rho^2) \leqslant 4M_2C_{\mathrm{LS}}\big(M_2^2C_{\mathrm{LS}}^2+1\big)H(\rho|\hat{\rho}). \qquad (7.31)
$$

Step 4: Control of Wasserstein and TV distances by $H(\rho|\hat{\rho})$. By inserting (7.31) into (7.29), and noting, by the definition of the Wasserstein distance,

$$
W_2^2(\rho^1, m_*) \leqslant \int W_2^2\big(\rho^{1|2}(\cdot|y), m_*\big)\rho^2(\mathrm{d}y),
$$

we get

$$
W_2^2(\rho^1, m_*) \leqslant \Big(2C_{\mathrm{LS}} + 4M_2C_{\mathrm{LS}}^2\big(M_2^2C_{\mathrm{LS}}^2+1\big)^{1/2}\Big)H(\rho|\hat{\rho}). \qquad (7.32)
$$

For the total variation distance, we observe that the Csiszár–Kullback–Pinsker inequality implies

$$
\begin{aligned}
&\int \big\|\rho^{1|2}(\cdot|y) - \hat{m}_y\big\|_{\mathrm{TV}}^2 \rho^2(\mathrm{d}y) \leqslant 2H(\rho|\hat{\rho}), \\
&\qquad \int \|\hat{m}_y - m_*\|_{\mathrm{TV}}^2\rho^2(\mathrm{d}y) \leqslant \int \big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y).
\end{aligned}
$$

By the triangle and Jensen's inequalities, we get

$$
\begin{aligned}
\|\rho^1 - m_*\|_{\mathrm{TV}}^2 &\leqslant \int \big\|\rho^{1|2}(\cdot|y) - m_*\big\|_{\mathrm{TV}}^2\rho^2(\mathrm{d}y) \\
&\leqslant 2\int \Big(\big\|\rho^{1|2}(\cdot|y) - \hat{m}_y\big\|_{\mathrm{TV}}^2 + \|\hat{m}_y - m_*\|_{\mathrm{TV}}^2\Big)\rho^2(\mathrm{d}y) \\
&\leqslant 4H(\rho|\hat{\rho}) + 2\int \big(H(\hat{m}_y|m_*) + H(m_*|\hat{m}_y)\big)\rho^2(\mathrm{d}y).
\end{aligned}
$$

Then, by inserting (7.31) into (7.28), we get

$$
\|\rho^1 - m_*\|_{\mathrm{TV}}^2 \leqslant \Big(4 + 8M_2C_{\mathrm{LS}}\big(M_2^2C_{\mathrm{LS}}^2+1\big)^{1/2}\Big)H(\rho|\hat{\rho}). \qquad (7.33)
$$

Step 5: Control of $H(\rho|\hat{\rho})$ and conclusion for the convex case. In the case where

$$M_1 \coloneqq \sup_{x\in\mathbb{R}^d, y\in\mathbb{R}^D} \ell(x)^\top \nabla^2\Phi(y)\ell(x) < \infty,$$

by Proposition 7.25 and (7.22) in Proposition 7.27, we immediately get

$$H(\rho|\hat{\rho}) \leqslant \frac{C_{\mathrm{LS}}}{2}(D + 2M_1)\lambda = H'.$$

In the general case where M_1 is not necessarily finite, Proposition 7.25 and (7.21) in Proposition 7.27 yield

$$H(\rho|\hat{\rho}) \leqslant \frac{C_{\mathrm{LS}}}{2}(D + 24M_2C_{\mathrm{LS}}d)\lambda + 24M_2C_{\mathrm{LS}}W_2^2(\rho^1, m_*)\lambda.$$

Together with the upper bound (7.32) of $W_2^2(\rho^1, m_*)$, we get

$$H(\rho|\hat{\rho}) \leqslant \frac{C_{\mathrm{LS}}(D + 24M_2C_{\mathrm{LS}}d)\lambda}{2 - 96M_2C_{\mathrm{LS}}^2\Big(1 + 2M_2C_{\mathrm{LS}}\big(M_2^2C_{\mathrm{LS}}^2 + 1\big)^{1/2}\Big)\lambda} = H,$$

for

$$\lambda < \frac{1}{48M_2C_{\mathrm{LS}}^2\Big(1 + 2M_2C_{\mathrm{LS}}\big(M_2^2C_{\mathrm{LS}}^2 + 1\big)^{1/2}\Big)} = \lambda_0.$$

We obtain the desired estimates on $v(\rho^2)$, Wasserstein and TV distances, by inserting the upper bounds of $H(\rho|\hat{\rho})$ for the respective cases into (7.31), (7.32), (7.33).

Now we work with a concave Φ and prove the last claim of the theorem.

Step 6: Case of concave Φ. Observe first that the mapping $y \mapsto \nabla^2\Phi(y)\cdot\hat{y}$ is a gradient:

$$\begin{aligned}\nabla^2\Phi(y)\cdot\hat{y} &= \nabla^2\Phi(y)\cdot\frac{\int \ell(x)\exp\big(-2\nabla\Phi(y)\cdot\ell(x)\big)\,\mathrm{d}x}{\int \exp\big(-2\nabla\Phi(y)\cdot\ell(x)\big)\,\mathrm{d}x}\\ &= -\frac{1}{2}\nabla_y\bigg(\log\int\exp\big(-2\nabla\Phi(y)\cdot\ell(x)\big)\,\mathrm{d}x\bigg).\end{aligned}$$

This identity is analogous to the fact in thermodynamics that when we derive the free energy with respect to a variable, we get the statistical average of its response variable. Moreover,

$$\nabla^2\Phi(y)\cdot y = \nabla_y\big(\nabla\Phi(y)\cdot y - \Phi(y)\big).$$

Thus, taking the test function

$$\varphi(y) = -\frac{1}{2}\log\int\exp\big(-2\nabla\Phi(y)\cdot\ell(x)\big)\,\mathrm{d}x - \nabla\Phi(y)\cdot y + \Phi(y)$$

in Lemma 7.24, we get

$$\iint \big(\ell(x) - y\big)^\top\nabla^2\Phi(y)(\hat{y} - y)\rho(\mathrm{d}x\,\mathrm{d}y) = 0.$$

Consequently,

$$
\begin{aligned}
-&\iint (\hat{y}-y)^\top \nabla^2\Phi(y)(\hat{y}-y)\rho(\mathrm{d}x\,\mathrm{d}y) \\
&= -\iint \big(\hat{y}-\ell(x)\big)^\top \nabla^2\Phi(y)(\hat{y}-y)\rho(\mathrm{d}x\,\mathrm{d}y) \\
&= \int \Big\langle \big(-\nabla^2\Phi(y)\big)^{1/2}\ell, \hat{m}_y - \rho^{1|2}(\cdot|y) \Big\rangle \cdot \big(-\nabla^2\Phi(y)\big)^{1/2}(\hat{y}-y)\rho^2(\mathrm{d}y) \\
&\leqslant \Big(-\int (\hat{y}-y)^\top \nabla^2\Phi(y)(\hat{y}-y)\rho^2(\mathrm{d}y)\Big)^{1/2} \\
&\qquad \times \Big(\int \Big|\Big\langle \big(-\nabla^2\Phi(y)\big)^{1/2}\ell, \hat{m}_y - \rho^{1|2}(\cdot|y) \Big\rangle\Big|^2 \rho^2(\mathrm{d}y)\Big)^{1/2},
\end{aligned}
$$

which implies that

$$
-\int (\hat{y}-y)^\top \nabla^2\Phi(y)(\hat{y}-y)\rho^2(\mathrm{d}y) \leqslant \int \Big|\Big\langle \big(-\nabla^2\Phi(y)\big)^{1/2}\ell, \hat{m}_y - \rho^{1|2}(\cdot|y) \Big\rangle\Big|^2 \rho^2(\mathrm{d}y).
$$

The term on the right satisfies

$$
\begin{aligned}
\int \Big|\Big\langle \big(-\nabla^2\Phi(y)\big)^{1/2}\ell, \hat{m}_y - \rho^{1|2}(\cdot|y) \Big\rangle\Big|^2 \rho^2(\mathrm{d}y) &\leqslant M_2 \int W_1^2\big(\hat{m}_y, \rho^{1|2}(\cdot|y)\big)\rho^2(\mathrm{d}y) \\
&\leqslant M_2 C_{\mathrm{LS}} \int H\big(\rho^{1|2}(\cdot|y)\big|\hat{m}_y\big)\rho^2(\mathrm{d}y)
\end{aligned}
$$

by the uniform LSI for $(\hat{m}_y)_{y\in\mathcal{K}}$. The entropy estimate in Proposition 7.25 gives

$$
\int H\big(\rho^{1|2}(\cdot|y)\big|\hat{m}_y\big)\rho^2(\mathrm{d}y) \leqslant \frac{C_{\mathrm{LS}}}{2}(D+2I)\lambda \leqslant \frac{C_{\mathrm{LS}} D\lambda}{2},
$$

as in the case of concave Φ, the term $I \leqslant 0$ by its definition (7.16). □

Chapter 8

Mean field optimization problem regularized by Fisher information

Abstract. Recently there is a rising interest in the research of mean field optimization, in particular because of its role in analyzing the training of neural networks. In this paper by adding the Fisher Information as the regularizer, we relate the regularized mean field optimization problem to a so-called mean field Schrödinger (MFS for short) dynamics. We develop an energy-dissipation method to show that the marginal distributions of the MFS dynamics converge exponentially quickly towards the unique minimizer of the regularized optimization problem. Remarkably, the MFS dynamics is proved to be a gradient flow on the probability measure space with respect to the relative entropy. Finally we propose a Monte Carlo method to sample the marginal distributions of the MFS dynamics.

Based on joint work with Julien Claisse, Giovanni Conforti and Zhenjie Ren.

8.1 Introduction

Recently the mean field optimization problem, namely

$$\inf_{p\in\mathcal{P}} \mathfrak{F}(p), \qquad \text{for a function } \mathfrak{F}:\mathcal{P}\to\mathbb{R}, \text{ where } \mathcal{P} \text{ is a set of probability measures,}$$

attracts increasing attention, in particular because of its role in analysing the training of artificial neural networks. The universal representation theorem (see e.g. [115]) ensures that a given continous function $f:\mathbb{R}^d\to\mathbb{R}$ can be approximated by the parametric form:

$$f(x)\approx\sum_{i=1}^{N} c_i\varphi(a_i\cdot x+b_i), \qquad \text{with } c_i\in\mathbb{R},\ a_i\in\mathbb{R}^d,\ b_i\in\mathbb{R} \text{ for } 1\leqslant i\leqslant N,$$

where φ is a fixed non-constant, bounded, continuous activation function. This particular parametrization is called a two-layer neural network (with one hidden layer). In order to train the optimal parameters, one needs to solve the optimization problem

$$\inf_{(c_i,a_i,b_i)_{1\leqslant i\leqslant N}} \sum_{j=1}^{M} L\Big(f(x_j), \sum_{i=1}^{N} c_i\varphi(a_i\cdot x_j+b_i)\Big),$$

where $L:(y,z)\mapsto L(y,z)$ is a loss function, typically convex in z. Here we face an overparametrized, non-convex optimization, and have no theory for an efficient solution. However it has been recently observed (see e.g. [162, 117, 57, 128]) that by lifting the optimization problem to the space of probability measures, namely

$$\inf_{p\in\mathcal{P}} \sum_{j=1}^{M} L\big(f(x_j), \mathbb{E}^p[C\varphi(A\cdot x+B)]\big),$$

with random variables (C,A,B) taking values in $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$ following the distribution p, one makes the optimization convex (notice that the function $F:p\mapsto \sum_{j=1}^{M} L\big(f(x_j), \mathbb{E}^p[C\varphi(A\cdot x+B)]\big)$ is convex), and has extensive tools to find the minimizers.

Unlike in [57] where the authors address the mean field optimization directly, in [162, 117] the authors add the entropy regularizer $H(p):=\int p(x)\log p(x)\,\mathrm{d}x$, that is, they aim at solving the regularized optimization problem:mfofisher-

$$\inf_{p\in\mathcal{P}} F(p)+\frac{\sigma^2}{2}H(p). \tag{8.1}$$

Recall the definition of the linear derivative $\delta F/\delta p$ and the intrinsic derivative D_pF (see Remark 8.2 below) for functions on the space of probability measures. In [117] the authors introduce the mean field Langevin (MFL for short) dynamics:

$$\mathrm{d}X_t = -D_pF(p_t,X_t)\,\mathrm{d}t+\sigma\,\mathrm{d}W_t,$$

where $p_t=\mathrm{Law}(X_t)$ and W is a standard Brownian motion, and prove that the marginal laws $(p_t)_{t\geqslant 0}$ of the MFL dynamics converge towards the minimizer of the entropic regularization (8.1). In the following works [178, 56] it has been shown that the convergence is exponentially quick.

In this paper we try to look into the mean field optimization problem from another perspective, by adding the Fisher information $I(p):=\int|\nabla\log p(x)|^2p(x)\,\mathrm{d}x$ instead of the entropy as the regularizer, namely solving the regularized optimization

$$\inf_{p\in\mathcal{P}}\mathfrak{F}^\sigma(p),\quad \mathfrak{F}^\sigma(p):=F(p)+\frac{\sigma^2}{4}I(p).$$

By convexity and calculus of variations (see Theorem 8.28), it is not hard to see that $p^*\in\mathrm{argmin}_{p\in\mathcal{P}}\mathfrak{F}^\sigma(p)$ if

$$\frac{\delta\mathfrak{F}^\sigma}{\delta p}(p^*,x):=\frac{\delta F}{\delta p}(p^*,x)-\frac{\sigma^2}{4}\big(2\Delta\log p^*+|\nabla\log p^*|^2\big)=\text{constant}. \tag{8.2}$$

We shall introduce the mean field Schrödinger (MFS for short) dynamics:

$$\partial_t p_t = -\frac{\delta\mathfrak{F}^\sigma}{\delta p}(p_t,\cdot)p_t,$$

prove its wellposedness and show that its marginal distributions $(p_t)_{t\geqslant 0}$ converges (uniformly) towards the minimizer of the free energy function $\mathfrak{F}^\sigma$. One crucial observation is that the free energy function decays along the MFS dynamics:

$$\frac{\mathrm{d}\mathfrak{F}^\sigma(p_t)}{\mathrm{d}t} = -\int \left| \frac{\delta \mathfrak{F}^\sigma}{\delta p}(p_t, x) \right|^2 p_t(\mathrm{d}x).$$

In order to prove it rigorously, we develop a probabilistic argument (coupling of diffusions) to estimate $(\nabla \log p_t, \nabla^2 \log p_t)_{t\geqslant 0}$. Remarkably, the estimate we obtain is uniform in time. Using the energy dissipation we can show that $(p_t)_{t\geqslant 0}$ converges exponentially quickly with help of the convexity of F and the Poincaré inequality. Another main contribution of this paper is to show that the MFS dynamics is a gradient flow of the free energy function $\mathfrak{F}^\sigma$ on the space of probability measures, provided that the 'distance' between the probability measures is measured by relative entropy. Finally it is noteworthy that MFS dynamics is numerically implementable, and we shall briefly propose a Monte Carlo simulation method.

Related works. Assume F to be linear, i.e. $F(p) := \int f(x)p(\mathrm{d}x)$ with a real potential function f and denote the wave function by $\psi := \sqrt{p}$. Then the function $\mathfrak{F}^\sigma$ reduces to the conventional energy function in quantum mechanics, composed of the potential energy $\langle \psi, f\psi \rangle_{L^2}$ and the kinetic energy $\sigma^2 \langle \nabla\psi, \nabla\psi \rangle_{L^2}$. Meanwhile, the MFS dynamics is reduced to the semigroup generated by the Schrödinger operator:

$$\partial_t \psi = -\mathcal{H}\psi, \quad \text{with } \mathcal{H} := -\frac{\sigma^2}{2}\Delta + \frac{1}{2}f. \tag{8.3}$$

The properties of the classical Schrödinger operator, including its longtime behavior, have been extensively studied in the literature, see e.g. the monographs [190, 151]. There are also profound studies in cases where F is nonlinear, notably the density functional theory [87, 88]. However, to our knowledge there is no literature dedicated to the category of convex potential $F : \mathcal{P} \to \mathbb{R}$, and studying the longtime behavior of such nonlinear Schrödinger operator by exploiting the convexity. In addition, the probabilistic nature of our arguments seems novel.

Using the change of variable: $u := -\log p^*$, the first order equation (8.2) can be rewritten as

$$\frac{\sigma^2}{2}\Delta u - \frac{\sigma^2}{4}|\nabla u|^2 + \frac{\delta F}{\delta p}(p^*, x) = \text{constant}.$$

So the function u solves an ergodic Hamilton-Jacobi-Bellman equation, and its gradient ∇u is the optimal control for the ergodic stochastic control problem:mfofisher-

$$\lim_{T\to\infty} \frac{1}{T} \sup_\alpha \mathbb{E}\left[\int_0^T \left(\frac{1}{2}|\alpha_t|^2 + \frac{2}{\sigma^2}\frac{\delta F}{\delta p}(p^*, X_t^\alpha) \right) \mathrm{d}t \right],$$

where

$$\mathrm{d}X_t^\alpha = \alpha_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

Further note that the probability $p^* = e^{-u}$ coincides with the invariant measure of the optimal controlled diffusion: $\mathrm{d}X_t^* = -\nabla u(X_t^*)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$, so that p^* is the Nash equilibrium of the corresponding ergodic mean field game. For more details on the ergodic mean field game, we refer to the seminal paper [143], and for more general mean field games we refer to the recent monographs [37, 38]. Our convergence result of the MFS dynamics $(p_t)_{t\geqslant 0}$ towards p^* offers an approximation to the equilibrium of the ergodic mean field game.

Our result on the gradient flow, as far as we know, is new to the literature. It is well known to the community of computational physics that the normalized solution $(\psi_t)_{t\geqslant 0}$ to the imaginary time Schrödinger equation (8.3) is the gradient flow of the free energy $\mathfrak{F}^\sigma$ on the L^2-unit ball. On the other hand, in [202] the authors discuss the (linear) optimization problem without Fisher information regularizer, and formally show that the dynamics, $\partial_t p = -fp$, is the gradient flow of the potential functional $\int f dp$ on the space of probability measures provided that the distance between the measures are measured by the relative entropy. Inspired by these works, we prove in the current paper that the solution to the variational problem

$$p^h_{i+1} := \operatorname*{argmin}_{p\in\mathcal{P}}\{\mathfrak{F}^\sigma(p) + h^{-1}H(p|p^h_i)\}, \qquad \text{for } h>0,\ i\geqslant 0,$$

converges to the continuous-time flow of the MFS dynamics as $h\to 0$. This result can be viewed as a counterpart of seminal paper [126] on the Wasserstein-2 gradient flow.

The rest of the paper is organized as follows. In Section 8.2 we formulate the problem and state the main results of the paper. The proofs are postponed to the subsequent sections. In Section 8.3, we show that the MFS dynamic is well-defined and admits an important decomposition as the exponential of a sum of a convex and a Lipschitz function. Then we study the long time behavior of this dynamic in Section 8.4 and we prove that it converges exponentially fast to the unique minimizer of the mean field optimization problem regularized by Fisher information. Finally we establish in Section 8.5 that the MFS dynamic corresponds to the gradient flow with respect to the relative entropy. Some technical results including a refined reflection coupling result are also gathered in the appendices.

Notations. (i) For each $T>0$, we denote by $Q_T = (0,T]\times\mathbb{R}^d$, $\bar{Q}_T = [0,T]\times\mathbb{R}^d$ and by $C^n(Q_T)$ the set of functions f such that $\partial_t^k\nabla^m f$ is continuous on Q_T for $2k+m\leqslant n$. In the case $T=+\infty$, we simply write $Q=(0,\infty)\times\mathbb{R}^d$, $\bar{Q} = [0,+\infty)\times\mathbb{R}^d$.

(ii) Given a measure μ on $\mathbb{R}^d$, let $W^{k,p}(\mu)$ be the Sobolev space of functions $f:\mathbb{R}^d\to\mathbb{R}$ such $f\in L^p(\mu)$ and $\nabla^l f\in L^p(\mu)$ for all $l\leqslant p$. In particular, we denote $H^1(\mu) := W^{1,2}(\mu)$. We simply write $W^{k,p}$ and H^1 when μ is the Lebesgue measure.

(iii) Let $\mathcal{P}_p(\mathbb{R}^d)$ be the collection of distribution on $\mathbb{R}^d$ with finite first p moments. It is equipped with $\mathcal{W}_p$ the Wasserstein distance of order p.

(iv) Given $u:\mathbb{R}^d\to\mathbb{R}$, we consider the functional norms

$$\|u\|_{(2)} := \sup_{x\in\mathbb{R}^d}\frac{|u(x)|}{1+|x|^2}, \quad \|u\|_\infty := \sup_{x\in\mathbb{R}^d}|u(x)|.$$

8.2 Main results

8.2.1 Free energy with Fisher information

Denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of all probability measures on $\mathbb{R}^d$ with finite second moments, endowed with $\mathcal{W}_2$ the Wasserstein distance of order 2. We focus on the probability measures admitting densities, and denote the density of $p \in \mathcal{P}_2(\mathbb{R}^d)$ still by $p : \mathbb{R}^d \to \mathbb{R}$ if it exists. In particular we are interested in the probability measures of density satifying:

$$\mathcal{P}_H := \{p \in \mathcal{P}_2(\mathbb{R}^d) : \sqrt{p} \in H^1\}.$$

In this paper we study a regularized mean field optimization problem, namely, given a potential function $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ we aim at solving

$$\inf_{p \in \mathcal{P}_H} \mathfrak{F}^\sigma(p), \qquad \text{with } \mathfrak{F}^\sigma(p) := F(p) + \sigma^2 I(p), \tag{8.4}$$

where $\sigma > 0$ and I is the Fisher information defined by

$$I(p) := \int_{\mathbb{R}^d} |\nabla \sqrt{p}(x)|^2 \, \mathrm{d}x. \tag{8.5}$$

In the literature, $\mathfrak{F}^\sigma$ is called the Ginzburg–Landau energy function with temperature σ. Note that for $p \in \mathcal{P}_H$ and $p > 0$, it holds

$$4 \int_{\mathbb{R}^d} |\nabla \sqrt{p}(x)|^2 \, \mathrm{d}x = \int_{\mathbb{R}^d} |\nabla \log p(x)|^2 p(x) \, \mathrm{d}x.$$

Throughout the paper, we assume that the potential function F is smooth, convex and coercive as stated in the following assumption.

Definition 8.1. We say that a function $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is $\mathcal{C}^1$ if there exist $\frac{\delta F}{\delta p} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ continuous with quadratic growth in the second variable such that for all p, $q \in \mathcal{P}_2(\mathbb{R}^d)$,

$$F(q) - F(p) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta p}\big(tq + (1-t)p, x\big)(q-p)(\mathrm{d}x) \, \mathrm{d}t.$$

Remark 8.2. Note that $F \in \mathcal{C}^1$ is $\mathcal{W}_2$-continuous and $\delta F/\delta p$ is defined up to constant. We call $\delta F/\delta p$ the linear derivative and we may further define the intrinsic derivative $D_p F(p,x) := \nabla \frac{\delta F}{\delta p}(p,x)$.

Assumption 8.3. Assume that F is $\mathcal{C}^1$, convex and

$$F(p) \geqslant \lambda \int_{\mathbb{R}^d} |x|^2 p(\mathrm{d}x) \qquad \text{for some } \lambda > 0.$$

The following proposition states that the bias caused by the regularizer vanishes as the temperature $\sigma \to 0$. It ensures that the Fisher information is efficient as regularizer in this mean field optimization problem.

Proposition 8.4. *It holds*

$$\lim_{\sigma \to 0} \inf_{p \in \mathcal{P}_H} \mathfrak{F}^\sigma(p) = \inf_{p \in \mathcal{P}_2} F(p).$$

Proof. Given $\varepsilon > 0$, let $p \in \mathcal{P}_2$ be such that $F(p) < \inf_{p \in \mathcal{P}_2} F(p) + \varepsilon$. By truncation and mollification, define $p_{K,\delta} := p_K * \varphi_\delta$ where $p_K := p \mathbb{1}_{|x| \leqslant K}/p(|x| \leqslant K)$ and $\varphi_\delta(x) := (2\pi\delta)^{-d/2} \exp(-|x|^2/2\delta)$. It is clear that $p_{K,\delta}$ converges to p in $\mathcal{W}_2$ as $K \to \infty$ and $\delta \to 0$. Additionally, one easily checks by direct computation that

$I(p_{K,\delta}) < +\infty$. By $\mathcal{W}_2$-continuity of F, we deduce by choosing K large and δ small enough that

$$\inf_{p\in\mathcal{P}_H} \mathfrak{F}^\sigma(p) \leqslant F(p_{K,\delta}) + \frac{\sigma^2}{2} I(p_{K,\delta}) \leqslant F(p) + \varepsilon + \frac{\sigma^2}{2} I(p_{K,\delta}) \leqslant \inf_{p\in\mathcal{P}_2} F(p) + 2\varepsilon + \frac{\sigma^2}{2} I(p_{K,\delta}).$$

We conclude by taking the limit $\sigma \to 0$. □

For the gradient flow analysis of Section 8.2.3 below, we shall actually consider a slightly more general mean field optimization problem. Namely, we aim at minimizing the following generalized free energy function: for all $p \in \mathcal{P}_H$,

$$\mathfrak{F}^{\sigma,\gamma}(p) := F(p) + \sigma^2 I(p) + \gamma H(p), \tag{8.6}$$

where $\gamma \geqslant 0$ and H is the entropy defined as

$$H(p) := \int_{\mathbb{R}^d} p(x) \log p(x)\, \mathrm{d}x.$$

By considering the limit of the rate of change $\big(\mathfrak{F}^{\sigma,\gamma}\big(p + t(q-p)\big) - \mathfrak{F}^{\sigma,\gamma}(p)\big)/t$ as $t \to 0$, a formal calculus leads to define by abuse of notation

$$\frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p,\cdot) := \frac{\delta F}{\delta p}(p,\cdot) - \frac{\sigma^2}{2}\Delta \log p - \frac{\sigma^2}{4}|\nabla \log p|^2 + \gamma \log p - \lambda(p), \tag{8.7}$$

where $\lambda(p) \in \mathbb{R}$ is chosen so that

$$\int_{\mathbb{R}^d} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p,x) p(x)\, \mathrm{d}x = 0. \tag{8.8}$$

The details of this calculation can be found within the proof of Theorem 8.28 below. Note also that equivalent formulas for $\delta\mathfrak{F}^{\sigma,\gamma}/\delta p$ can be obtained by observing that

$$\Delta \log p + \frac{1}{2}|\nabla \log p|^2 = \frac{\Delta p}{p} - \frac{1}{2}\frac{|\nabla p|^2}{p^2} = 2\frac{\Delta\sqrt{p}}{\sqrt{p}}.$$

8.2.2 Mean field Schrödinger dynamics

Given the definition in (8.7), we will consider the following generalized mean field Schrödinger (MFS for short) dynamics

$$\partial_t p_t = -\frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t,\cdot) p_t.$$

Thanks to the normalization in (8.8), the mass of p_t is conserved to 1. Writing the functional derivative explicitly, we have the following dynamics

$$\partial_t p_t = -\Big(\frac{\delta F}{\delta p}(p_t,\cdot) - \frac{\sigma^2}{2}\Delta \log p_t - \frac{\sigma^2}{4}|\nabla \log p_t|^2 + \gamma \log p_t - \lambda_t\Big) p_t \tag{8.9}$$

where $p_t = p(t,\cdot)$ and $\lambda_t = \lambda(p_t)$ satisfies

$$\lambda_t = \int_{\mathbb{R}^d}\Big(\frac{\delta F}{\delta p}(p_t,x) - \frac{\sigma^2}{2}\Delta \log p_t(x) - \frac{\sigma^2}{4}|\nabla \log p_t(x)|^2 + \gamma \log p_t(x)\Big) p_t(x)\, \mathrm{d}x.$$

In particular, the important case $\gamma = 0$ is called the MFS dynamics, namely,

$$\partial_t p_t = -\frac{\delta \mathfrak{F}^\sigma}{\delta p}(p_t, \cdot) p_t. \tag{8.10}$$

Intuitively the generalized MFS dynamics follows the direction of steepest descent as it moves in the opposite direction of the derivative $\delta \mathfrak{F}^{\sigma,\gamma}/\delta p$. To ensure that it is indeed converging towards a minimizer of $\mathfrak{F}^{\sigma,\gamma}$, the crucial assumption in this paper is that the derivative $\delta F/\delta p$ decomposes into the sum of a convex potential and a Lipschitz perturbation as stated below.

Assumption 8.5. The linear derivative admits the decomposition $\frac{\delta F}{\delta p}(p, x) = g(x) + G(p, x)$ where g and $G(p, \cdot)$ are C^2 such that

(i) g is $\underline{\kappa}$-convex and has bounded Hessian, i.e.,

$$\underline{\kappa} I_d \leqslant \nabla^2 g \leqslant \bar{\kappa} I_d, \qquad \text{for some } \bar{\kappa} \geqslant \underline{\kappa}.$$

(ii) G is $\mathcal{W}_1$-continuous in p and Lipschitz continuous in x, i.e., for all x, $y \in \mathbb{R}^d$, $p \in \mathcal{P}_2(\mathbb{R}^d)$,

$$|G(p, x) - G(p, y)| \leqslant L_G |x - y|.$$

(iii) ∇G is Lipschitz continuous, i.e., for all x, $y \in \mathbb{R}^d$, p, $q \in \mathcal{P}_2(\mathbb{R}^d)$,

$$|\nabla G(p, x) - \nabla G(q, y)| \leqslant L_G\big(|x - y| + \mathcal{W}_1(p, q)\big).$$

Assumption 8.6. The initial distribution admits the decomposition

$$p_0(x) = \exp\big(-v_0(x) - w_0(x)\big),$$

where v_0 and w_0 are C^1 such that

(i) v_0 is $\underline{\eta}_0$-convex and ∇v_0 is Lipschitz continuous, i.e., for all x, $y \in \mathbb{R}^d$,

$$|\nabla v_0(x) - \nabla v_0(y)| \leqslant \bar{\eta}_0 |x - y|, \quad \big(\nabla v_0(x) - \nabla v_0(y)\big) \cdot (x - y) \geqslant \underline{\eta}_0 |x - y|^2.$$

(ii) w_0 and ∇w_0 are both Lipschitz continuous, i.e., for all x, $y \in \mathbb{R}^d$,

$$|w_0(x) - w_0(y)| + |\nabla w_0(x) - \nabla w_0(y)| \leqslant L_0 |x - y|.$$

In the sequel, we assume that Assumptions 8.3, 8.5 and 8.6 hold. First we show that the generalized MFS dynamic is well-defined and that it decomposes as the exponential of a sum of a convex and a Lipschitz function. The proof is postponed to Section 8.3.2.

Theorem 8.7. *Under the assumptions above, the generalized MFS dynamics* (8.9) *admits a unique positive classical solution* $p \in C^3(Q) \cap C(\bar{Q})$. *In addition, it admits the decomposition* $p_t = \exp(-v_t - w_t)$ *where there exist* $\underline{\eta}$, $\bar{\eta}$, $L > 0$ *such that*

$$\underline{\eta} I_d \leqslant \nabla^2 v_t \leqslant \bar{\eta} I_d, \quad \|\nabla w_t\|_\infty \vee \|\nabla^2 w_t\|_\infty \leqslant L, \qquad \forall t > 0. \tag{8.11}$$

Then we study the long-time behaviour of the generalized MFS dynamics and establish convergence toward the unique minimizer of the generalized free energy function. The proof is postponed to Section 8.4.3. It essentially relies on energy dissipation which can be derived formally as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{F}^{\sigma,\gamma}(p_t) = \int_{\mathbb{R}^d} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t, x)\partial_t p_t(x)\,\mathrm{d}x = -\int_{\mathbb{R}^d} \left|\frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t, x)\right|^2 p_t(x)\,\mathrm{d}x,$$

See Theorem 8.29 below for a proof. It follows that the generalized free energy monotonously decreases along the generalized MFS dynamics (8.9). Intuitively, the dissipation of energy only stops at the moment $\delta\mathfrak{F}^{\sigma,\gamma}/\delta p\,(p^*, \cdot) = 0$. Since $\mathfrak{F}^{\sigma,\gamma}$ is (strictly) convex, it is a sufficient condition for p^* to be the minimizer, see Theorem 8.28 below.

Theorem 8.8. *Under the assumptions above, the solution $(p_t)_{t\geqslant 0}$ to (8.9) converges uniformly on $\mathbb{R}^d$ to p^*, the unique minimizer of $\mathfrak{F}^{\sigma,\gamma}$ in $\mathcal{P}_H$. In addition, the optimizer p^* satisfies (8.11) and it is a stationary solution to (8.9), i.e.,*

$$\frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p^*, \cdot) = 0. \tag{8.12}$$

Remark 8.9. By Lemma 8.32 below, the family of distributions $(p_t)_{t\geqslant 0}$ admits uniform Gaussian bounds and thus it also converges to p^* for the L^p-norm or the $\mathcal{W}_p$-distance for any $p \geqslant 1$.

Remark 8.10. In case that the function $p \mapsto F(p)$ is linear, i.e., $F(p) = \int_{\mathbb{R}^d} f(x)p(\mathrm{d}x)$ with some potential f, the function $\mathfrak{F}^\sigma$ is the classical energy function in quantum mechanics composed of the potential energy F and the kinetic one $\int_{\mathbb{R}^d} |\nabla\sqrt{p}(x)|^2\,\mathrm{d}x$. Let p^* be the minimizer of $\mathfrak{F}^\sigma$, and denote by $\psi^* := \sqrt{p^*}$ the corresponding wave function. Then the first order equation (8.12) reads

$$-\sigma^2\Delta\psi^* + f\psi^* = c\psi^*, \qquad \text{with } c = \mathfrak{F}^\sigma(p^*) = \min_{p\in\mathcal{P}_H} \mathfrak{F}^\sigma(p).$$

It is well known that c is the smallest eigenvalue of the Schrödinger operator $-\sigma^2\Delta + f$ and that ψ^* is the ground state of the quantum system.

Further we shall prove that the convergence for the MFS dynamics (with $\gamma = 0$) is exponentially quick. See Section 8.4.4 below for a proof. As a byproduct, we establish a functional inequality in Theorem 8.34 which may carry independent interest.

Theorem 8.11. *There exists a constant $c(\underline{\eta}, \bar{\eta}, L, d, \sigma) > 0$ such that*

$$\mathfrak{F}^\sigma(p_t) - \mathfrak{F}^\sigma(p^*) \leqslant e^{-ct}\big(\mathfrak{F}^\sigma(p_0) - \mathfrak{F}^\sigma(p^*)\big). \tag{8.13}$$

Moreover, it holds

$$\frac{\sigma^2}{4} I(p_t|p^*) \leqslant e^{-ct}\big(\mathfrak{F}^\sigma(p_0) - \mathfrak{F}^\sigma(p^*)\big),$$

where $I(p_t|p^*) := \int p_t|\nabla\log(p_t/p^*)|^2$ is the relative Fisher information.

8.2.3 Gradient flow with relative entropy

In this paper, we shall further investigate the gradient flow of the free energy function $\mathfrak{F}^\sigma$ with respect to the relative entropy. First, given $h > 0$ and a distribution $\tilde{p}$ satisfying Assumption 8.6, consider the variational problem:mfofisher-

$$\inf_{p \in \mathcal{P}_H} \{\mathfrak{F}^\sigma(p) + h^{-1} H(p|\tilde{p})\}. \tag{8.14}$$

where $H(p|\tilde{p}) := \int p \log(p/\tilde{p})$ is the relative entropy. In view of Assumption 8.6, we have the decomposition $\tilde{p} = \exp(-\tilde{u})$ with $\tilde{u} = \tilde{v} + \tilde{w}$. Denoting by

$$\tilde{F}(p) := F(p) + h^{-1} \int_{\mathbb{R}^d} \tilde{u}(x) p(\mathrm{d}x)$$

the new potential function, we may rewrite the objective function in the optimization (8.14) in the form of the generalized free energy function (8.6), i.e.

$$\tilde{\mathfrak{F}}^{\sigma,h^{-1}}(p) = \tilde{F}(p) + \sigma^2 I(p) + h^{-1} H(p).$$

Moreover, the new potential function $\tilde{F}$ still satisfies Assumption 8.5 with $\tilde{g} = g + h^{-1}\tilde{v}$ and $\tilde{G} = G + h^{-1}\tilde{w}$. Therefore, the following result is a straightforward consequence of Theorem 8.8.

Corollary 8.12. *If $\tilde{p}$ satisfies Assumption 8.6, the minimization problem (8.14) admits a unique minimizer $p^* \in \mathcal{P}_H$ still satisfying Assumption 8.6 (with different coefficients) and it satisfies the first order condition*

$$\frac{\delta \tilde{\mathfrak{F}}^{\sigma,h^{-1}}}{\delta p}(p^*, \cdot) = 0.$$

Now given $p_0^h := p_0$ satisfying Assumption 8.6, we may define a sequence of probability measures using the variational problem (8.14):

$$p_i^h := \operatorname*{argmin}_{p \in \mathcal{P}_H} \{\mathfrak{F}^\sigma(p) + h^{-1} H(p|p_{i-1}^h)\}, \qquad \text{for } i \geqslant 1. \tag{8.15}$$

It corresponds to the so-called *minimizing movement scheme* in the optimal transport literature. According to Corollary 8.12, the minimizer p_i^h is well defined and it satisfies the first order condition:

$$\frac{\delta \mathfrak{F}^\sigma}{\delta p}(p_i^h, \cdot) + h^{-1}(\log p_i^h - \log p_{i-1}^h) = \int h^{-1}(\log p_i^h - \log p_{i-1}^h) p_i^h. \tag{8.16}$$

Thus we expect as $h \to 0$ that the minimizing movement scheme p^h converges to the corresponding gradient flow p satisfying

$$\frac{\delta \mathfrak{F}^\sigma}{\delta p}(p_t, \cdot) + \partial_t \log p_t = 0,$$

which corresponds to the MFS dynamics (8.10).

This result is proved rigorously in Section 8.5.3 below. By slightly abusing the notations, define the continuous-time flow of probability measures:

$$p_t^h := p_i^h, \qquad \text{for } t \in [hi, h(i+1)).$$

Theorem 8.13. *The sequence of functions $(p^h)_{h>0}$ converges, uniformly on $[0,T]\times\mathbb{R}^d$ for any $T>0$, to p the MFS dynamics (8.10).*

Remark 8.14. In view of Corollary 8.37 below, the family of distributions $(p^h)_{h>0}$ admits uniform Gaussian bounds and thus we also have for any $p\geqslant 1$,

$$\sup_{t\in[0,T]}\left\|p_t^h-p_t\right\|_{L^p}\xrightarrow[h\to 0]{}0,\qquad \sup_{t\in[0,T]}\mathcal{W}_p\big(p_t^h,p_t\big)\xrightarrow[h\to 0]{}0.$$

8.2.4 Numerical simulation

In this section we shall briefly report how to sample $N^{-1}\sum_{i=1}^N\delta_{X_t^i}$ to approximate the probability law p_t in the MFS dynamics (8.10), without pursuing mathematical rigorism.

Observe first that the MFS dynamics (8.10) can be rewritten as

$$\partial_t p_t=\frac{\sigma^2}{2}\Delta p_t-\left(\frac{\delta F}{\delta p}(p_t,\cdot)+\frac{\sigma^2}{4}|\nabla\log p_t|^2-\lambda_t\right)p_t.$$

This can viewed as the Fokker–Planck equation describing the marginal distribution of a Brownian motion $(X_t)_{t\geqslant 0}$ killed at rate $\eta(t,x)\coloneqq\delta F/\delta p\,(p_t,x)+\sigma^2|\nabla\log p_t(x)|^2/4$ conditionned on not being killed. In other words, the particle X moves freely in the space $\mathbb{R}^d$ as a Brownian motion $(\sigma W_t)_{t\geqslant 0}$ before it gets killed with conditional probability

$$\mathbb{P}\big[X\text{ gets killed in }[t,t+\Delta t]\;\big|\;X_t\big]\approx\eta(t,X_t)\Delta t,\qquad\text{for small }\Delta t.$$

Meanwhile the killed particle gets reborn instantaneously according to the distribution p_t. This interpretation of the MFS dynamics offers an insight on how to sample the marginal law p_t. However, in order to evaluate the death rate $\eta(t,X_t)$, one needs to evaluate $|\nabla\log p_t|^2$, which can be hard if not impossible in practice. This difficulty forces us to find a more sophisticated way to sample p_t.

Now observe that $\psi_t\coloneqq\sqrt{p_t}$ solves the PDE:

$$\partial_t\psi_t=\frac{\sigma^2}{2}\Delta\psi_t-\frac{1}{2}\left(\frac{\delta F}{\delta p}(p_t,\cdot)-\lambda_t\right)\psi_t.\tag{8.17}$$

Then introduce two scalings of ψ_t, namely, $\bar\psi_t\coloneqq\exp\big(-\frac12\int_0^t\lambda_s\,\mathrm{d}s\big)\psi_t$ and $\hat\psi_t\coloneqq\psi_t/\int\psi_t$ so that

$$\partial_t\bar\psi_t=\frac{\sigma^2}{2}\Delta\bar\psi_t-\frac{1}{2}\frac{\delta F}{\delta p}(p_t,\cdot)\bar\psi_t,\qquad\partial_t\hat\psi_t=\frac{\sigma^2}{2}\Delta\hat\psi_t-\frac{1}{2}\left(\frac{\delta F}{\delta p}(p_t,\cdot)-\hat\lambda_t\right)\hat\psi_t,$$

where the constant $\hat\lambda_t\in\mathbb{R}$ is chosen so that $\hat\psi_t$ is a probability density. Observe that:

- By the Feynman–Kac formula, the function $\bar\psi$ has the probabilistic representation:

$$\begin{aligned}\bar\psi_t(x)&=\mathbb{E}\left[\exp\left(-\int_0^t\frac{1}{2}\frac{\delta F}{\delta p}(p_{t-s},x+\sigma W_s)\,\mathrm{d}s\right)\psi_0(x+\sigma W_t)\right]\\&\approx\frac{1}{M}\sum_{j=1}^M\exp\left(-\int_0^t\frac{1}{2}\frac{\delta F}{\delta p}(p_{t-s},x+\sigma W_s^j)\,\mathrm{d}s\right)\psi_0(x+\sigma W_t^j),\end{aligned}$$

  where the latter is the standard Monte Carlo approximation of the expectation.

- The probability law $\hat{\psi}$ is the marginal distribution of a Brownian motion killed at rate $\eta(t,x) := \frac{1}{2}\frac{\delta F}{\delta p}(p_t, x)$ conditioned on not being killed. It can be sampled by simulating a large number $(\hat{X}^i)_{1\leqslant i\leqslant N}$ of independent Brownian particles killed at rate η which upon dying are instantaneously reborn by duplicating one of the living particles.
- Eventually, the distribution p_t can be approximately sampled as the following weighted empirical measure

$$p_t = \frac{\bar{\psi}_t}{\int_{\mathbb{R}^d} \bar{\psi}_t(x)\hat{\psi}_t(x)\,\mathrm{d}x}\hat{\psi}_t \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\bar{\psi}(t, \hat{X}_t^i)}{\frac{1}{N}\sum_{k=1}^{N} \bar{\psi}(t, \hat{X}_t^k)}\delta_{\hat{X}_t^i}.$$

Remark 8.15. In particular, in view of Remark 8.10, the Monte Carlo method above offers an efficient way to sample the ground state of a high dimensional quantum system. To our knowledge there is little discussion on similar numerical schemes in the literature.

8.3 Mean field Schrödinger dynamics

In order to study the generalized MFS dynamics in (8.9), we introduce a change of variable $p_t = \exp(-u_t)/\int \exp(-u_t)$ where u satisfies the following equation:

$$\partial_t u_t = \frac{\sigma^2}{2}\Delta u_t - \frac{\sigma^2}{4}|\nabla u_t|^2 + \frac{\delta F}{\delta p}(p_t, \cdot) - \gamma u_t, \tag{8.18}$$

with initial condition $u_0 = -\log p_0$. Clearly, u is a classical solution to (8.18) if and only if the probability density p is a positive classical solution to (8.9). Thus we consider the mapping

$$(m_t)_{t\in[0,T]} \mapsto (u_t)_{t\in[0,T]} \mapsto (p_t)_{t\in[0,T]} \tag{8.19}$$

where $p_t = \exp(-u_t)/\int \exp(-u_t)$ and u solves the equation

$$\partial_t u_t = \frac{\sigma^2}{2}\Delta u_t - \frac{\sigma^2}{4}|\nabla u_t|^2 + \frac{\delta F}{\delta p}(m_t, \cdot) - \gamma u_t, \tag{8.20}$$

and we look for a fixed point to this mapping as it corresponds to a solution to (8.18). Note that (8.20) corresponds to the Hamilton–Jacobi–Bellman (HJB for short) equation of a classical linear-quadratic stochastic control problem and so u is well-defined as the unique viscosity solution of this equation by standard arguments.

In this section, we first show that the solution to the HJB equation (8.20) can be decomposed as the sum of a convex and a Lipschitz function. This allows us to apply a reflection coupling argument to show that the mapping (8.19) is a contraction on short horizon and thus to ensure existence and uniqueness of the solution to (8.18).This completes the proof of Theorem 8.7. Finally we gather some properties of the solution to (8.9) for later use.

8.3.1 Hamilton–Jacobi–Bellman equation

The aim of this section is to prove that the solution to the HJB equation (8.20) is smooth and can be decomposed into the sum of a convex and a Lipschitz function as stated in Proposition 8.17 below. Throughout this section we assume that the following assumption holds.

Assumption 8.16. Assume that the mapping $t \mapsto m_t$ is $\mathcal{W}_1$-continuous, i.e.,

$$\lim_{s\to t} \mathcal{W}_1(m_t, m_s) = 0, \qquad \text{for all } t \geqslant 0.$$

Proposition 8.17. *There exists a unique classical solution $u \in C^3(Q) \cap C(\bar{Q})$ to the HJB equation (8.20). In addition, $u = v + w$ where there exist $\underline{\eta}$, $\bar{\eta}$, $L > 0$, independent of m, such that*

$$\underline{\eta} I_d \leqslant \nabla^2 v_t \leqslant \bar{\eta} I_d, \quad \|\nabla w_t\|_\infty \vee \|\nabla^2 w_t\|_\infty \leqslant L, \qquad \forall t > 0.$$

By the Cole–Hopf transformation, we may prove in a rather classical way that there exists a unique smooth solution to (8.20). We refer to Appendix E.1 for a complete proof. Further, given the decomposition $\delta F/\delta p\,(p,x) = g(x) + G(p,x)$ in Assumption 8.5 and $u_0 = v_0 + w_0$ in Assumption 8.6, we are tempted to decompose the solution to (8.20) as $u = v + w$, where v solves the HJB equation corresponding to the convex part

$$\partial_t v_t = \frac{\sigma^2}{2}\Delta v_t - \frac{\sigma^2}{4}|\nabla v_t|^2 + g - \gamma v_t, \tag{8.21}$$

and w solves the remaining part

$$\partial_t w_t = \frac{\sigma^2}{2}\Delta w_t - \frac{\sigma^2}{2}\nabla v_t \cdot \nabla w_t - \frac{\sigma^2}{4}|\nabla w_t|^2 + G(m_t,\cdot) - \gamma w_t. \tag{8.22}$$

Because it is a special case of (8.20), (8.21) also admits a unique classical solution, and therefore so does (8.22). The proof of Proposition 8.17 is completed through Propositions 8.20, 8.21 and 8.22 below.

Remark 8.18. In case $G = 0$ and $w_0 = 0$, we have $u = v$. Therefore all the properties proved for the function u are shared by the function v.

Lemma 8.19. *Let u be the classical solution to (8.20). There exists a constant $\delta > 0$ only depending on $\bar{\kappa}$, $\bar{\eta}_0$, L_0, L_G from Assumption 8.5 and 8.6 such that $\sup_{T\leqslant\delta} \|\nabla^2 u(T,\cdot)\|_\infty < \infty$.*

Proof. Step 1. We first show that the SDE (8.25) below admits a unique strong solution. Define $\psi(t,x) := \exp\big(-u(t,x)/2\big)$. By Lemma E.2 in the appendices, we have

$$\psi(t,x) = \mathbb{E}\bigg[\exp\bigg(-\frac{1}{2}\int_0^t \Big(\frac{\delta F}{\delta p}(m_{t-s}, x+\sigma W_s) - \gamma u(t-s, x+\sigma W_s)\Big)\,\mathrm{d}s\bigg)\psi_0(x+\sigma W_t)\bigg]. \tag{8.23}$$

Now consider the continuous paths space $C([0,T])$ as the canonical space. Denote by $(\bar{\mathcal{F}}_t)_{t\leqslant T}$ the canonical filtration and $\bar{X}$ the canonical process. Let $\mathbb{P}$ be the probability measure such that $(\bar{X} - x)/\sigma$ is a $\mathbb{P}$-Brownian motion starting from the origin. We may define an equivalent probability measure $\mathbb{Q}$ on the canonical space via

$$\frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\bar{\mathcal{F}}_T} = \Lambda_T := \exp\bigg(-\int_0^T \frac{1}{2}\Big(\frac{\delta F}{\delta p}(m_{T-s}, \bar{X}_s) - \gamma u(t-s, \bar{X}_s)\Big)\,\mathrm{d}t\bigg)\psi_0(\bar{X}_T)\Big/\psi(T,x). \tag{8.24}$$

By Itô's formula, we may identify that

$$
\begin{aligned}
&\mathbb{E}^{\mathbb{P}}[\Lambda_T|\bar{\mathcal{F}}_t] \\
&\quad = \exp\Big(-\int_0^t \frac{1}{2}\Big(\frac{\delta F}{\delta p}(m_{T-s}, \bar{X}_s) - \gamma u(t-s, \bar{X}_s)\Big)\,\mathrm{d}t\Big)\psi(T-t, \bar{X}_t)\Big/\psi(T,x) \\
&\quad = \exp\Big(-\int_0^t \frac{1}{2}\nabla u(t-s, \bar{X}_s)\cdot \mathrm{d}\bar{X}_s - \int_0^t \frac{\sigma^2}{8}|\nabla u(t-s, \bar{X}_s)|^2\,\mathrm{d}s\Big).
\end{aligned}
$$

Using the Girsanov's theorem, we may conclude that the SDE

$$
X_t = x - \int_0^t \frac{\sigma^2}{2}\nabla u(T-s, X_s)\,\mathrm{d}s + \sigma W_t, \tag{8.25}
$$

admits a weak solution. In addition, since $x \mapsto \nabla u(t,x)$ is locally Lipschitz, the SDE above has the property of pathwise uniqueness. Therefore, we can conclude by the Yamada–Watanabe theorem.

Step 2. Next we observe that ∇u is the classical solution to

$$
\partial_t \nabla u_t = \frac{\sigma^2}{2}\Delta \nabla u_t - \frac{\sigma^2}{2}\nabla^2 u_t \nabla u_t + \nabla \frac{\delta F}{\delta p}(m_t, \cdot) - \gamma \nabla u_t. \tag{8.26}
$$

By denoting $Y_t \coloneqq \nabla u(T-t, X_t)$, it follows from Itô's formula that (X,Y) solves the forward-backward SDE (FBSDE for short):

$$
\begin{cases}
\mathrm{d}X_t = -\frac{\sigma^2}{2}Y_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, & X_0 = x, \\
\mathrm{d}Y_t = \big(\gamma Y_t - \nabla \frac{\delta F}{\delta p}(m_{T-t}, X_t)\big)\,\mathrm{d}t + Z_t\,\mathrm{d}W_t, & Y_T = \nabla u_0(X_T),
\end{cases}
$$

where $Z_t = \sigma \nabla^2 u(T-t, X_t)$. Introduce the norm

$$
\|(Y,Z)\|_{\mathcal{D}} \coloneqq \sup_{t\leqslant T}\Big\{\mathbb{E}\Big[|Y_t|^2 + \int_t^T |Z_s|^2 ds\Big]\Big\}^{1/2}.
$$

We are going to show that $\|(Y,Z)\|_{\mathcal{D}} < \infty$, provided that T is small enough.

By Lemma E.1 and Proposition E.3 in the appendices, we have

$$
\exp\big(-C_T(1+|x|^2)\big) \leqslant \psi(t,x) \leqslant C_T, \qquad |\nabla\psi(t,x)| \leqslant C_T(1+|x|^2).
$$

Therefore,

$$
|\nabla u(t,x)| = 2\frac{|\nabla \psi|}{\psi}(t,x) \leqslant C_T(1+|x|^2)\exp(C_T|x|^2).
$$

On the other hand, by the definition of Λ_T in (8.24), we have

$$
\Lambda_T \leqslant C_T \exp\Big(C_T\big(|x|^2 + \sup_{t\leqslant T}|\bar{X}_t|^2\big)\Big).
$$

Now we may provide the following estimate

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{t\leqslant T}|Y_t|^2\Big] &= \mathbb{E}\Big[\sup_{t\leqslant T}|\nabla u(T-t, X_t)|^2\Big] = \mathbb{E}^{\mathbb{P}}\Big[\Lambda_T \sup_{t\leqslant T}|\nabla u(T-t, \bar{X}_t)|^2\Big] \\
&\leqslant C_T e^{C_T|x|^2}\mathbb{E}^{\mathbb{P}}\Big[\big(1+\sup_{t\leqslant T}|\bar{X}_t|^2\big)\exp\big(C_T\sup_{t\leqslant T}|\bar{X}_t|^2\big)\Big].
\end{aligned}
$$

In particular, if T is small enough, we have

$$\mathbb{E}^{\mathbb{P}}\big[\big(1+\sup_{t\leqslant T}|\bar{X}_t|^2\big)\exp\big(C_T\sup_{t\leqslant T}|\bar{X}_t|^2\big)\big]<\infty.$$

Moreover, by Itô's formula, we obtain

$$\begin{aligned}\mathrm{d}|Y_t|^2 &= \left(2\gamma|Y_t|^2 - 2Y_t\cdot\nabla\frac{\delta F}{\delta p}(m_{T-t},X_t)+|Z_t|^2\right)\mathrm{d}t + 2Y_t\cdot Z_t\,\mathrm{d}W_t\\ &\geqslant \left((2\gamma-1)|Y_t|^2 - \left|\nabla\frac{\delta F}{\delta p}(m_{T-t},X_t)\right|^2+|Z_t|^2\right)\mathrm{d}t + 2Y_t\cdot Z_t\,\mathrm{d}W_t.\end{aligned}$$

Define the stopping time $\tau_n \coloneqq \inf\{t\geqslant 0 : |Z_t|\geqslant n\}$, and note that

$$\begin{aligned}\mathbb{E}\left[\int_0^{T\wedge\tau_n}|Z_t|^2\,\mathrm{d}t\right] &\leqslant \mathbb{E}[|Y_{T\wedge\tau_n}|^2]-\mathbb{E}[|Y_0|^2]\\ &\qquad + \mathbb{E}\left[\int_0^{T\wedge\tau_n}\left((1-2\gamma)|Y_t|^2+\left|\nabla\frac{\delta F}{\delta p}(m_{T-t},X_t)\right|^2\mathrm{d}t\right)\right].\end{aligned}$$

Since we have proved $\mathbb{E}\big[\sup_{t\leqslant T}|Y_t|^2\big]<\infty$, by monotone and dominated convergence theorem, we obtain

$$\mathbb{E}\left[\int_0^T|Z_t|^2\,\mathrm{d}t\right]\leqslant\mathbb{E}\big[|Y_T|^2\big]+\mathbb{E}\left[\int_0^T\left((1-2\gamma)|Y_t|^2+\left|\nabla\frac{\delta F}{\delta p}(m_{T-t},X_t)\right|^2\mathrm{d}t\right)\right]<\infty.$$

Therefore, we have $\|(Y,Z)\|_{\mathcal{D}}<\infty$.

Step 3. It is known (see e.g. [158, Theorem I.5.1]) that there exists $\delta>0$ only depending on $\bar{\kappa}$, $\bar{\eta}_0$, L_0, L_G such that for $T\leqslant\delta$ the process (Y,Z) here is the unique solution to the FBSDE such that $\|(Y,Z)\|_{\mathcal{D}}<\infty$. Moreover, by standard a priori estimate (again see [158, Theorem I.5.1]) we may find a constant $C\geqslant 0$ only depending on $\bar{\kappa}$, $\bar{\eta}_0$, L_0, L_G such that for (Y',Z') solution to the FBSDE above starting from $X_0=x'$ we have

$$\|(Y,Z)-(Y',Z')\|_{\mathcal{D}}\leqslant C|x-x'|,\qquad\text{for } T\leqslant\delta.$$

In particular, it implies that

$$|Y_0-Y_0'|=|\nabla u(T,x)-\nabla u(T,x')|\leqslant C|x-x'|,$$

so that $\sup_{T\leqslant\delta}\|\nabla^2 u(T,\cdot)\|_\infty<\infty$. □

Proposition 8.20. *Let v be the classical solution to (8.21). It holds:*

(i) *The function v_t is η_t-convex, i.e., $\nabla^2 v_t\geqslant\eta_t I_d$, with*

$$\frac{d\eta_t}{dt}=\underline{\kappa}-\gamma\eta_t-\sigma^2\eta_t^2,\qquad \eta_0=\underline{\eta}_0. \tag{8.27}$$

In particular, v_t is $\underline{\eta}$-convex with $\underline{\eta}\coloneqq\min\big(\eta_0,(\sqrt{\gamma^2+4\sigma^2\underline{\kappa}}-\gamma)/(2\sigma^2)\big)$.

(ii) *The Hessian of v is bounded uniformly w.r.t. $t>0$ and m.*

Proof. We divide the following discussion into three steps.

Step 1. We first prove the strict convexity of the solution v on a short horizon. Fix $T := \delta$ small enough so that, thanks to Lemma 8.19, $\nabla^2 v$ is uniformly bounded on $(0,T]$. We shall prove that not only $\nabla^2 v$ has a positive lower bound, but also the bound does not depend on T.

As in Step 1 of the proof of Lemma 8.19, we may define the strong solution

$$X_t = x - \int_0^t \frac{\sigma^2}{2}\nabla v(T-s, X_s)\,\mathrm{d}s + \sigma W_t.$$

Further define $Y_t := \nabla v(T-t, X_t)$ and $Z_t := \sigma\nabla^2 v(T-t, X_t)$ so that $\|(Y,Z)\|_{\mathcal{D}} < \infty$ and (Y,Z) is the unique solution to the FBSDE on the short horizon $[0,T]$:

$$\begin{cases} \mathrm{d}X_t = -\frac{\sigma^2}{2}Y_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, & X_0 = x, \\ \mathrm{d}Y_t = \big(\gamma Y_t - \nabla g(X_t)\big)\,\mathrm{d}t + Z_t\,\mathrm{d}W_t, & Y_T = \nabla v_0(X_T). \end{cases}$$

Define (X', Y', Z') similarly with $X'_0 = x'$, and further denote by $\delta X_t := X_t - X'_t$, $\delta Y_t := Y_t - Y'_t$, $\delta Z_t := Z_t - Z'_t$. Note that due to the uniqueness of the solution to the FBSDE, we have $\delta X_t = \delta Y_t = \delta Z_t = 0$ for $t \geqslant \tau := \inf\{t \geqslant 0 : \ \delta X_t = 0\}$. By Itô's formula, it is easy to verify that

$$\begin{aligned} &\mathrm{d}\frac{\delta X_t \cdot \delta Y_t}{|\delta X_t|^2} \\ &\quad = \bigg(-\frac{\sigma^2|\delta Y_t|^2}{2|\delta X_t|^2} + \gamma\frac{\delta X_t\cdot\delta Y_t}{|\delta X_t|^2} - \frac{\delta X_t\cdot\big(\nabla g(X_t) - \nabla g(X'_t)\big)}{|\delta X_t|^2} + \sigma^2\frac{|\delta X_t\cdot\delta Y_t|^2}{|\delta X_t|^4}\bigg)\,\mathrm{d}t \\ &\qquad + \frac{\delta X_t\cdot\delta Z_t\,\mathrm{d}W_t}{|\delta X_t|^2}. \end{aligned}$$

Therefore, the pair $(\hat{Y}_t, \hat{Z}_t) := (\delta X_t\cdot\delta Y_t/|\delta X_t|^2, \delta X_t^\top\delta Z_t/|\delta X_t|^2)$ solves the BSDE:

$$\mathrm{d}\hat{Y}_t = \bigg(-\frac{\sigma^2|\delta Y_t|^2}{2|\delta X_t|^2} + \gamma\hat{Y}_t - \frac{\delta X_t\cdot\big(\nabla g(X_t) - \nabla g(X'_t)\big)}{|\delta X_t|^2} + \sigma^2\hat{Y}_t^2\bigg)\,\mathrm{d}t + \hat{Z}_t\,\mathrm{d}W_t.$$

According to Lemma 8.19, the process $\hat{Y}$ is bounded on $[0,T]$ and so is the coefficient in front of $\mathrm{d}t$ above. By the Itô isometry, we clearly have $\mathbb{E}\big[\int_0^T |\hat{Z}_t|^2\,\mathrm{d}t\big] < \infty$.

We aim at providing a lower bound for $\hat{Y}$. Consider the Riccati equation (8.27) and note that the solution $(\eta_t)_{t\geqslant 0}$ evolves monotonously from the initial condition $\eta_0 > 0$ to the positive equilibrium $\eta^* := (\sqrt{\gamma^2 + 4\sigma^2\underline{\kappa}} - \gamma)/(2\sigma^2)$. In particular, it holds

$$\underline{\eta} = \min(\eta_0, \eta^*) \leqslant \eta_t \leqslant \max(\eta_0, \eta^*). \tag{8.28}$$

Define $\hat{\eta}_t := \eta_{T-t}$ for $t \leqslant T$ so that

$$\mathrm{d}\hat{\eta}_t = (-\underline{\kappa} + \gamma\hat{\eta}_t + \sigma^2\hat{\eta}_t^2)\,\mathrm{d}t, \qquad \hat{\eta}_T \leqslant \hat{Y}_T.$$

Since g is $\underline{\kappa}$-convex, we have

$$\begin{aligned} &\mathrm{d}(\hat{Y}_t - \hat{\eta}_t) \\ &\quad = \bigg(-\frac{\sigma^2|\delta Y_t|^2}{2|\delta X_t|^2} - \frac{\delta X_t\cdot\big(\nabla g(X_t) - \nabla g(X'_t)\big)}{|\delta X_t|^2} + \underline{\kappa} + \gamma(\hat{Y}_t - \hat{\eta}_t) + \sigma^2(\hat{Y}_t^2 - \hat{\eta}_t^2)\bigg)\,\mathrm{d}t \\ &\qquad + \hat{Z}_t\,\mathrm{d}W_t \\ &\quad \leqslant \big(\gamma(\hat{Y}_t - \hat{\eta}_t) + \sigma^2(\hat{Y}_t + \hat{\eta}_t)(\hat{Y}_t - \hat{\eta}_t)\big)\,\mathrm{d}t + \hat{Z}_t\,\mathrm{d}W_t \end{aligned}$$

Since $\hat{Y}_t$, $\hat{\eta}_t$ are both bounded and $\mathbb{E}\big[\int_0^T |\hat{Z}_t|^2\,\mathrm{d}t\big] < \infty$, it follows from the standard comparison principle for BSDE that $\hat{Y}_t - \hat{\eta}_t \geqslant 0$, i.e., the function v_t is η_t-convex for $t \in [0, T]$.

Step 2. We shall improve the bound of $|\nabla^2 v|$ to get a bound independent of the horizon $T = \delta$. Note that ∇v satisfies the equation

$$\partial_t \nabla v_t = \frac{\sigma^2}{2}\Delta \nabla v_t - \frac{\sigma^2}{2}\nabla^2 v_t \nabla v_t + \nabla g - \gamma \nabla v_t.$$

Thus it admits the probabilistic representation

$$\nabla v(t,x) = \mathbb{E}\bigg[\int_0^t e^{-\gamma s}\nabla g(X_s)\,\mathrm{d}s + e^{-\gamma t}\nabla v_0(X_t)\bigg],$$

with

$$X_s = x - \int_0^s \frac{\sigma^2}{2}\nabla v(t-r, X_r)\,\mathrm{d}r + \sigma W_s.$$

Let X' be the solution to the SDE above with $X'_0 = x'$. Since ∇g and ∇v_0 are both Lipschitz continuous, we have

$$|\nabla v(t,x) - \nabla v(t,x')| \leqslant \mathbb{E}\bigg[\bar{\kappa}\int_0^t |X_s - X'_s|\,\mathrm{d}s + \bar{\eta}_0 |X_t - X'_t|\bigg], \tag{8.29}$$

Now recall that we have proved in Step (i) that the function v_s is η_s-convex for $s \in [0, t]$ so that

$$\begin{aligned}
\frac{1}{2}\,\mathrm{d}|X_s - X'_s|^2 &= (X_s - X'_s)\cdot(\mathrm{d}X_s - \mathrm{d}X'_s) \\
&= -\frac{\sigma^2}{2}(X_s - X'_s)\cdot\big(\nabla v(t-s, X_s) - \nabla v(t-s, X'_s)\big)\,\mathrm{d}s \\
&\leqslant -\frac{\sigma^2 \eta_{t-s}}{2}|X_s - X'_s|^2\,\mathrm{d}s,
\end{aligned}$$

Furthermore recall that $\eta_s \geqslant \underline{\eta}$ for all $s \geqslant 0$ by (8.28) so that

$$|X_s - X'_s| \leqslant \exp(-\sigma^2 \underline{\eta} s/2)|x - x'|. \tag{8.30}$$

Together with (8.29), we obtain

$$|\nabla v(t,x) - \nabla v(t,x')| \leqslant C\bigg(1 + \frac{2}{\sigma^2 \underline{\eta}}\bigg)|x - x'|.$$

Therefore $|\nabla^2 v(t,\cdot)| \leqslant C\big(1 + 2/(\sigma^2\underline{\eta})\big)$, in particular the bound does not depend on $T = \delta$.

Step 3. By the result of Step 2, we know that $\nabla^2 v(\delta,\cdot)$ is bounded and the bound does not depend on δ. Together with Lemma 8.19, we conclude that $\nabla^2 v$ is bounded on $[\delta, 2\delta]$, and further deduce that v_t is η_t-convex and $\nabla^2 v$ has a δ-independent bound again on $[\delta, 2\delta]$ thanks to the results of Steps 1 and 2. Therefore the desired result follows from induction. □

Proposition 8.21. *Let w be the classical solution to (8.22). Then the function $x \mapsto w(t,x)$ is Lipschitz continuous uniformly w.r.t. $t \geqslant 0$ and m.*

Proof. We consider the following stochastic control problem. Let $(\Omega, \mathcal{F}, \mathbb{P}, \mathbb{F})$ be a filtered probability space, and W be a $(\mathbb{P},\mathbb{F})$-Brownian motion. Denote by $\mathcal{A}$ the collection of admissible control process, i.e., α is progressively measurable and $\mathbb{E}\big[\int_0^t |\alpha_t|^2\,\mathrm{d}t\big] < \infty$. Then it follows from standard dynamic programming arguments that

$$w(t,x) = \inf_{\alpha\in\mathcal{A}} \mathbb{E}\bigg[\int_0^t e^{-\gamma s}\Big(G(m_{t-s}, X_s^\alpha) + \frac{\sigma^2}{4}|\alpha_s|^2\Big)\,\mathrm{d}s + e^{-\gamma t} w_0(X_t^\alpha)\bigg],$$

where X^α stands for the strong solution to

$$\mathrm{d}X_s^\alpha = -\frac{\sigma^2}{2}\big(\nabla v(t-s, X_s^\alpha) + \alpha_s\big)\,\mathrm{d}s + \sigma\,\mathrm{d}W_s, \qquad X_0^\alpha = x.$$

Denote by Y^α the solution to the SDE above with $Y_0^\alpha = y$. Then it holds

$$|w(t,y) - w(t,x)| \leqslant \sup_\alpha \mathbb{E}\bigg[L_G\int_0^t e^{-\gamma s}|Y_s^\alpha - X_s^\alpha|\,\mathrm{d}s + L_0 e^{-\gamma t}|Y_t^\alpha - X_t^\alpha|\bigg]. \quad (8.31)$$

Using the convexity of v_s from Proposition 8.20, we obtain by the same argument as (8.30) that

$$|Y_s^\alpha - X_s^\alpha| \leqslant \exp(-\sigma^2 \underline{\eta} s/2)|y-x|.$$

Together with (8.31), we can find a (t,m)-independent constant $L>0$ such that

$$|w(t,y) - w(t,x)| \leqslant L|y-x|. \qquad \square$$

Given the decomposition of u as the sum of a convex and a Lipschitz function, we shall also prove that the Hessian of u is bounded uniformly in time which is clearly an improvement over Lemma 8.19.

Proposition 8.22. *Let u be the classical solution to (8.20). Then the Hessian of u is bounded uniformly w.r.t. $t>0$ and m.*

Proof. Recall that ∇u satisfies (8.26) so that, by Feynman–Kac's formula, it admits the probabilistic representation

$$\nabla u(t,x) = \mathbb{E}\bigg[\int_0^t e^{-\gamma s}\nabla\frac{\delta F}{\delta p}(m_{t-s}, X_s)\,\mathrm{d}s + e^{-\gamma t}\nabla u_0(X_t)\bigg], \quad (8.32)$$

with

$$X_s = x - \frac{\sigma^2}{2}\int_0^s \nabla u(t-r, X_r)\,\mathrm{d}r + \sigma W_s.$$

Let us prove that $x \mapsto \nabla u(t,x)$ is Lipschitz continuous with a Lipschitz constant independent of t and m. Denote by Y the solution to the SDE above with $Y_0 = y$. It follows from the reflection coupling Theorem E.7 in the appendices that for $p_s^X := \mathcal{L}(X_s)$ and $p_s^Y := \mathcal{L}(Y_s)$,

$$\mathcal{W}_1(p_s^X, p_s^Y) \leqslant C\exp(-c\sigma^2 s)\mathcal{W}_1(p_0^X, p_0^Y), \qquad \text{for all } s \geqslant 0.$$

Note that the drift $\nabla u = \nabla v + \nabla w$ satisfies Assumption E.5 since v is $\underline{\eta}$-convex and ∇w is bounded, see Remark E.6. Together with (8.32) and the fact that $\nabla \frac{\delta F}{\delta p}(p, \cdot)$ and ∇u_0 are uniformly Lipschitz, we have by Kantorovich duality that

$$|\nabla u(t,x) - \nabla u(t,y)| \leqslant C\bigg(\int_0^t \mathcal{W}_1(p_s^X, p_s^Y)\,\mathrm{d}s + \mathcal{W}_1(p_t^X, p_t^Y)\bigg) \leqslant C|x-y|,$$

where the constant C does not depend on t and m. □

8.3.2 Proof of Theorem 8.7

Proof of Theorem 8.7. In view of Proposition 8.17, it is enough to show that the mapping (8.19) $(m_t)_{t\in[0,T]} \mapsto (p_t)_{t\in[0,T]}$ is a contraction for T small enough, where $p_t = \exp(-u_t)/\int \exp(-u_t)$ with u the solution to (8.20). This contraction property relies essentially on a reflection coupling argument established in Appendix E.3 which follows from the decomposition of u as the sum of a convex and a Lipschtz function.

Step 1. Let $(\tilde{m}_t)_{t\in[0,T]}$ be another flow of probability measures satisfying Assumption 8.16, and use it to define the function $\tilde{u}$ as in (8.20). Denote by $\delta u := u - \tilde{u}$. Using the stability result for the HJB equation (8.20) proved in Proposition 8.23 below, we obtain

$$\sup_{t\leqslant T} \|\nabla \delta u(t,\cdot)\|_\infty \leqslant TC_T \sup_{t\leqslant T} \mathcal{W}_1(m_t, \tilde{m}_t). \tag{8.33}$$

Step 2. Further define the probability density $\tilde{p}_t = \exp(-\tilde{u}_t)/\int \exp(-\tilde{u}_t)$. Note that p_t and $\tilde{p}_t$ are the invariant measures of the diffusion processes

$$\mathrm{d}X_s = -\nabla u(t, X_s)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W_s, \quad \mathrm{d}\tilde{X}_s = -\nabla \tilde{u}(t, \tilde{X}_s)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W_s,$$

respectively. Denote by $p_{t,s} := \mathcal{L}(X_s)$ and $\tilde{p}_{t,s} := \mathcal{L}(\tilde{X}_s)$ the marginal distributions, and assume that $p_{t,0} = \tilde{p}_{t,0} = p_0$. By Proposition 8.17 and Remark E.6, we may apply the reflection coupling in Theorem E.7 in the appendices to obtain

$$\mathcal{W}_1\big(p_{t,s}, \tilde{p}_{t,s}\big) \leqslant Ce^{-cs} \int_0^s e^{cr} \|\nabla \delta u(t,\cdot)\|_\infty \,\mathrm{d}r.$$

Let $s \to \infty$ on both sides. Since $\lim_{s\to\infty} \mathcal{W}_1(p_{t,s}, p_t) = 0$ and $\lim_{s\to\infty} \mathcal{W}_1(\tilde{p}_{t,s}, \tilde{p}_t) = 0$ by Remark E.8, we deduce that

$$\mathcal{W}_1\big(p_t, \tilde{p}_t\big) \leqslant C\|\nabla \delta u(t,\cdot)\|_\infty.$$

Step 2. Together with (8.33), we finally obtain

$$\sup_{t\leqslant T} \mathcal{W}_1\big(p_t, \tilde{p}_t\big) \leqslant TC_T \sup_{t\leqslant T} \mathcal{W}_1\big(m_t, \tilde{m}_t\big).$$

Therefore, given T small enough, the mapping $(m_t)_{t\leqslant T} \mapsto (p_t)_{t\leqslant T}$ is a contraction under the metric $\sup_{t\leqslant T} \mathcal{W}_1(\cdot_t, \cdot_t)$. □

The following lemma shows that the gradient ∇u of the solution to the HJB equation (8.20) is stable with respect to $(m_t)_{t\in[0,T]}$ as needed for the proof of Theorem 8.7 above, as well as with respect to ∇u_0 for later use.

Lemma 8.23. *Let $\tilde{u}$ be the classical solution to (8.20) corresponding to the flow of distribution $\tilde{m}$ satisfying Assumption 8.16 and the initial value $\tilde{u}_0$ satisfying Assumption 8.6. Then we have the following stability results:*

(i) *If $\nabla u_0 = \nabla \tilde{u}_0$, then $\|\nabla \delta u(t,\cdot)\|_\infty \leqslant C_t \int_0^t \mathcal{W}_1(m_s, \tilde{m}_s)\,\mathrm{d}s$.*

(ii) *Otherwise $\|\nabla \delta u(t,\cdot)\|_{(2)} \leqslant C_t\big(\int_0^t \mathcal{W}_1(m_s,\tilde{m}_s)\,\mathrm{d}s + \|\nabla\delta u_0\|_{(2)}\big)$.*

Proof. Similiar to (8.32), it follows from the Feynman-Kac's formula that

$$\nabla u(t,x) = \mathbb{E}\bigg[\int_0^t e^{-\gamma s}\nabla \frac{\delta F}{\delta p}(m_{t-s}, X_s)\,\mathrm{d}s + e^{-\gamma t}\nabla u_0(X_t)\bigg],$$
$$\nabla \tilde{u}(t,x) = \mathbb{E}\bigg[\int_0^t e^{-\gamma s}\nabla \frac{\delta F}{\delta p}(\tilde{m}_{t-s}, \tilde{X}_s)\,\mathrm{d}s + e^{-\gamma t}\nabla \tilde{u}_0(\tilde{X}_t)\bigg],$$

with

$$\mathrm{d}X_s = -\frac{\sigma^2}{2}\nabla u(t-s, X_s)\,\mathrm{d}s + \sigma\,\mathrm{d}W_s, \qquad X_0 = x,$$
$$\mathrm{d}\tilde{X}_s = -\frac{\sigma^2}{2}\nabla \tilde{u}(t-s, \tilde{X}_s)\,\mathrm{d}s + \sigma\,\mathrm{d}W_s, \qquad \tilde{X}_0 = x.$$

By Proposition 8.17 and Remark E.6, we may apply the reflection coupling in Theorem E.7 in the appendices to compare the marginal distribution of X and $\tilde{X}$, denoted by p and $\tilde{p}$ respectively. We obtain

$$\mathcal{W}_1(p_s, \tilde{p}_s) \leqslant C e^{-c\sigma^2 s}\int_0^s e^{c\sigma^2 r}\mathbb{E}\big[|\nabla\delta u(t-r, X_r)|\big]\,\mathrm{d}r.$$

Further, by Kantorovich duality and Lipschitz continuity of $\nabla\frac{\delta F}{\delta p}$ and $\nabla\tilde{u}_0$, we have

$$\begin{aligned}|\nabla\delta u(t,x)| \leqslant C\mathbb{E}\bigg[&\int_0^t\!\!\int_0^s Ce^{-\gamma s - c\sigma^2(s-r)}\big|\nabla\delta u(t-r, X_r)\big|\,\mathrm{d}r\,\mathrm{d}s \\ &+ \int_0^t e^{-\gamma s}\mathcal{W}_1(m_{t-s}, \tilde{m}_{t-s})\,\mathrm{d}s \\ &+ \int_0^t Ce^{-\gamma t - c\sigma^2(t-s)}|\nabla\delta u(t-s, X_s)|\,\mathrm{d}s + e^{-\gamma t}|\nabla\delta u_0(X_t)|\bigg],\end{aligned}$$

which implies that

$$|\nabla\delta u(t,x)| \leqslant C\mathbb{E}\bigg[\int_0^t \big|\nabla\delta u(t-s, X_s)\big|\,\mathrm{d}s + \int_0^t \mathcal{W}_1(m_s,\tilde{m}_s)ds + |\nabla\delta u_0(X_t)|\bigg]. \tag{8.34}$$

Recall the decomposition of the solution established in Proposition 8.17: $u = v + w$, $\tilde{u} = \tilde{v} + \tilde{w}$, where v, $\tilde{v}$ are strictly convex and w, $\tilde{w}$ are Lipschitz. We divide the following discussion into two cases.

Case 1. We assume $\nabla\delta u_0 = 0$. Note that in this case $\nabla v = \nabla\tilde{v}$ (because v, $\tilde{v}$ are not influenced by m or $\tilde{m}$) and that $\nabla\delta u = \nabla w - \nabla\tilde{w}$ is bounded. It follows from the (8.34) that

$$\|\nabla\delta u(t,\cdot)\|_\infty \leqslant C\bigg(\int_0^t \|\nabla\delta u(s,\cdot)\|_\infty\,\mathrm{d}s + \int_0^t \mathcal{W}_1(m_s,\tilde{m}_s)\,\mathrm{d}s\bigg).$$

Finally, by the Grönwall inequality, we obtain

$$\|\nabla\delta u(t,\cdot)\|_\infty \leqslant C_t\int_0^t \mathcal{W}_1(m_s,\tilde m_s)\,\mathrm{d}s.$$

Case 2. We consider the general case. Recall that both ∇v and $\nabla\tilde v$ are Lipschitz, and both ∇w and $\nabla\tilde w$ are bounded, so we have $\|\nabla\delta u(t,\cdot)\|_{(2)} < \infty$. Further it follows from (8.34) that

$$\begin{aligned}|\nabla\delta u(t,x)| &\leqslant C\bigg(\int_0^t \|\nabla\delta u(t-s,\cdot)\|_{(2)}\big(1+\mathbb{E}\big[|X_s|^2\big]\big)\,\mathrm{d}s\\ &\qquad\qquad + \int_0^t \mathcal{W}_1(m_s,\tilde m_s)\,\mathrm{d}s + \|\nabla\delta u_0\|_{(2)}\big(1+\mathbb{E}\big[|X_t|^2\big]\big)\bigg)\\ &\leqslant Ce^{ct}\bigg(\int_0^t \|\nabla\delta u(t-s,\cdot)\|_{(2)}(1+|x|^2)\,\mathrm{d}s\\ &\qquad\qquad + \int_0^t \mathcal{W}_1(m_s,\tilde m_s)\,\mathrm{d}s + \|\nabla\delta u_0\|_{(2)}(1+|x|^2)\bigg).\end{aligned}$$

Finally, by the Grönwall inequality, we obtain

$$\|\nabla\delta u(t,\cdot)\|_{(2)} \leqslant C_t\bigg(\int_0^t \mathcal{W}_1(m_s,\tilde m_s)\,\mathrm{d}s + \|\nabla\delta u_0\|_{(2)}\bigg). \qquad \square$$

8.3.3 Properties of mean field Schrödinger dynamics

The decomposition of the generalized MFS dynamics provided by Theorem 8.7 allows us to derive Gaussian bounds, first locally in time as stated below and later uniformly in time, see Lemma 8.32.

Proposition 8.24. *For any $T>0$, there exist $\underline{c}$, $\bar c$, $\underline{C}$, $\bar C > 0$, such that for all $t\in[0,T]$, $x\in\mathbb{R}^d$,*

$$\underline{C}\exp(-\underline{c}|x|^2) \leqslant p_t(x) \leqslant \bar C\exp(-\bar c|x|^2).$$

In particular, $p_t\in\mathcal{P}_H$ for all $t\geqslant 0$.

Proof. The Gaussian bounds follow immediately from Lemma E.4 in the appendices, whose assumptions are satisfied on $\mathcal{T}=[0,T]$ according to Theorem 8.7. Then we observe

$$|\nabla\sqrt{p_t}|^2 = \frac{1}{4}|\nabla\log p_t|^2 p_t \leqslant C_T(1+|x|^2)p_t,$$

where the latter follows from the boundedness of $\nabla^2\log p_t$. Thus $\nabla\sqrt{p_t}\in L^2$ and $p_t\in\mathcal{P}_H$. $\square$

Then we establish a stability result for the generalized MFS dynamics (8.9). It plays a crucial role in the proof of convergence in Theorem 8.8.

Proposition 8.25. *For $n\in\mathbb{N}$, let p^n (resp. p) be the generalized MFS dynamics (8.9) starting from p_0^n (resp. p_0), where p_0^n (resp. p_0) satisfy Assumption 8.6. If $\nabla\log p_0^n$ converges to $\nabla\log p_0$ in $\|\cdot\|_{(2)}$, then $(p_t^n,\log p_t^n)$ converges to $(p_t,\nabla\log p_t)$ in $\mathcal{W}_1\otimes\|\cdot\|_{(2)}$ for all $t>0$.*

Proof. Recall that the function u_t solution to (8.18) differs from $-\log p_t$ only through an additive constant (depending on t), in particular $\nabla u_t = -\nabla \log p_t$. Denote by $\delta u := u^n - u$. By the stability result of the HJB equation (8.20) proved in Proposition 8.23, we have

$$\|\nabla \delta u(T,\cdot)\|_{(2)} \leqslant C_T\Big(\int_0^T \mathcal{W}_1(p_t^n, p_t)\,\mathrm{d}t + \|\nabla \delta u_0\|_{(2)}\Big). \tag{8.35}$$

As in the proof of Theorem 8.7, note that p_t^n and p_t are the invariant measures of the diffusions:

$$\mathrm{d}X_s^n = -\nabla u^n(t, X_s^n)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W_s, \quad \mathrm{d}X_s = -\nabla u(t, X_s)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W_s,$$

respectively. Denote the marginal distributions $p_{t,s}^n := \mathcal{L}(X_s^n)$ and $p_{t,s} := \mathcal{L}(X_s)$, and assume that $p_{t,0}^n = p_{t,0}$. Using the reflection coupling, we deduce from Theorem E.7 that

$$\mathcal{W}_1(p_{t,s}^n, p_{t,s}) \leqslant Ce^{-cs}\int_0^s e^{cr}\mathbb{E}[|\nabla \delta u(t, X_r)|]\,\mathrm{d}r.$$

By letting $s \to \infty$ on both sides, it follows from using successively the $\mathcal{W}_1$-convergence of $p_{t,s}^n$ and $p_{t,s}$ toward p_t^n and p_t by Remark E.8, the linear growth of $\nabla \delta u(t,\cdot)$ and Lemma 8.24 that

$$\mathcal{W}_1(p_t^n, p_t) \leqslant C\int_{\mathbb{R}^d} |\nabla \delta u(t,x)| p_t(x)\,\mathrm{d}x \leqslant C_T\|\nabla \delta u(t,\cdot)\|_{(2)}.$$

Together with (8.35), by the Grönwall inequality, we obtain

$$\|\nabla \delta u(T,\cdot)\|_{(2)} \leqslant C_T e^{TC_T}\|\nabla \delta u_0\|_{(2)},$$

as well as

$$\mathcal{W}_1(p_T^n, p_T) \leqslant C_T e^{TC_T}\|\nabla \delta u_0\|_{(2)}. \qquad \square$$

8.4 Convergence towards the minimizer

8.4.1 First order condition

The aim of this section is to derive a first order condition to characterize the minimizer of the generalized free energy $\mathcal{F}^{\sigma,\gamma}$. Recall that $\mathcal{F}^{\sigma,\gamma}(p) = F(p) + \sigma^2 I(p) + \gamma H(p)$ with parameters $\sigma > 0$, $\gamma \geqslant 0$, and $I(p) = \int |\nabla\sqrt{p}|^2$, $H(p) = \int p\log p$.

Proposition 8.26. *The function $\mathfrak{F}^{\sigma,\gamma}$ is convex on $\mathcal{P}_H$. Additionally, if it admits a minimizer $p^* \in \mathcal{P}_H$ such that $1/p^* \in L^\infty_{loc}$, then it is unique.*

Proof. It follows from the convexity of F by Assumption 8.3, the convexity of H by convexity of $x \mapsto x\log x$ and Proposition 8.27 below. $\square$

Lemma 8.27. *Let p, $q \in \mathcal{P}_H$ and α, $\beta > 0$. Then we have*

$$I(\alpha p + \beta q) \leqslant \alpha I(p) + \beta I(q).$$

If in addition $1/p \in L^\infty_{loc}$, then the equality holds if and only if $p = q$.

Proof. Let $\varphi = \sqrt{p}$, $\psi = \sqrt{q}$. We have by using the Cauchy–Schwarz inequality

$$\begin{aligned} I(\alpha p + \beta q) &= \int \big|\nabla\sqrt{\alpha\varphi^2 + \beta\psi^2}\big|^2 = \int \frac{(\alpha\varphi\nabla\varphi + \beta\psi\nabla\psi)^2}{\alpha\varphi^2+\beta\psi^2} \\ &\leqslant \int \frac{(\alpha\varphi^2+\beta\psi^2)\big(\alpha(\nabla\varphi)^2 + \beta(\nabla\psi)^2\big)}{\alpha\varphi^2+\beta\psi^2} = \alpha I(p) + \beta I(q). \end{aligned}$$

The equality holds if and only if $\varphi\nabla\psi = \psi\nabla\varphi$. If in addition $1/p \in L^\infty_{\mathrm{loc}}$ then $1/\varphi \in L^\infty_{\mathrm{loc}}$ and $\psi/\varphi \in L^1_{\mathrm{loc}}$ which is a distribution in the sense of Schwartz. Its derivative satisfies

$$\nabla\left(\frac{\psi}{\varphi}\right) = \frac{\varphi\nabla\psi - \psi\nabla\varphi}{\varphi^2} = 0.$$

Therefore ψ/φ is constant a.e., i.e., p and q are proportional. □

Proposition 8.28. *If a probability measure $p \in \mathcal{P}_H$ satisfies $p \in C^2$ and*

$$p(x) \leqslant Ce^{-c|x|^2}, \quad |\nabla^2 \log p(x)| \leqslant C,$$

then the following inequality holds: for all $q \in \mathcal{P}_H$,

$$\mathfrak{F}^{\sigma,\gamma}(q) - \mathfrak{F}^{\sigma,\gamma}(p) \geqslant \int_{\mathbb{R}^d} \frac{\delta\mathfrak{F}^{\sigma,\gamma}}{\delta p}(p,x)\big(q(x)-p(x)\big)\,\mathrm{d}x.$$

In particular, if $\delta\mathfrak{F}^{\sigma,\gamma}/\delta p\,(p,\cdot) = 0$, then p is the unique minimizer of the generalized free energy $\mathfrak{F}^{\sigma,\gamma}$.

Proof. We have $\mathfrak{F}^{\sigma,\gamma}(p) = F(p) + \sigma^2 I(p) + \gamma H(p)$. We deal with each of these three terms separately. Adding the three subsequent inequalities gives the desired inequality. The second assertion then follows immediately from Theorem 8.26. Throughout the proof, we denote $p_t \coloneqq p + t(q-p)$ for $t \in [0,1]$.

Step 1. By convexity of F, it holds

$$F(q) - F(p) \geqslant \frac{F(p_t) - F(p)}{t}.$$

Since F is $\mathcal{C}^1$, we conclude by passing to the limit $t \to 0$ that

$$F(q) - F(p) \geqslant \frac{\mathrm{d}F(p_t)}{\mathrm{d}t}\bigg|_{t=0^+} = \int_{\mathbb{R}^d} \frac{\delta F}{\delta p}(p,\cdot)(q-p).$$

Step 2. Denote $I_K(p) \coloneqq \int_K |\nabla p|^2/4p$ for $K \subset \mathbb{R}^d$ compact and $p \in \mathcal{P}_H$. Assume first that q is bounded and compactly supported. Then it follows from the convexity of I_K and differentiation under the integral sign that

$$I_K(q) - I_K(p) \geqslant \frac{\mathrm{d}I_K(p_t)}{\mathrm{d}t}\bigg|_{t=0^+} = -\frac{1}{4}\int_K \frac{|\nabla p|^2}{p^2}(q-p) + \frac{1}{2}\int_K \frac{\nabla p \cdot \nabla(q-p)}{p}.$$

Note that $\nabla q = 2\sqrt{q}\,\nabla\sqrt{q} \in L^2$. Next we take the limit $K \uparrow \mathbb{R}^d$ and we observe that the r.h.s. converges by using for the first term, $|\nabla p(x)|/p(x) = |\nabla \log p(x)| \leqslant C(1+|x|)$ and p, $q \in \mathcal{P}_2$, and for the second term, $|\nabla p|^2/p = 4|\nabla\sqrt{p}|^2 \in L^1$. Using further integration by parts, since $p(x) \leqslant Ce^{-c|x|^2}$ and q is compactly supported, we obtain

$$I(q) - I(p) \geqslant -\frac{1}{4}\int_{\mathbb{R}^d} \left(\frac{|\nabla p|^2}{p^2} + 2\nabla \cdot \left(\frac{\nabla p}{p}\right)\right)(q-p).$$

To conclude it remains to deal with the general case $q \in \mathcal{P}_H$ not necessarily bounded and compactly supported. Given $M > 0$, we consider the distribution $q_M \propto \mathbb{1}_{|x|\leqslant M}\, q \wedge M$ and we apply the inequality above to q_M. Taking the limit $M \to \infty$ yields the desired result as the r.h.s. converges since $q \in \mathcal{P}_2$ and $|\nabla^2 \log p| \leqslant C$.

Step 3. Denote $H_K(p) \coloneqq \int_K p \log p$ for $K \subset \mathbb{R}^d$ compact and $p \in \mathcal{P}_H$. Assume first that q is bounded. Then it follows from the convexity of H_K and differentiation under the integral sign that

$$H_K(q) - H_K(p) \geqslant \left.\frac{\mathrm{d}H_K(p_t)}{\mathrm{d}t}\right|_{t=0^+} = \int_K (1+\log p)(q-p).$$

Next we take the limit $K \uparrow \mathbb{R}^d$ and we observe that the r.h.s. converges as p, $q \in \mathcal{P}_2$ and $|\log p(x)| \leqslant C(1+|x|^2)$. We obtain

$$H(q) - H(p) \geqslant \int_{\mathbb{R}^d} (\log p)(q-p)$$

To conclude it remains to deal with the general case $q \in \mathcal{P}_H$ not necessarily bounded. Given $M > 0$, we consider the distribution $q_M \propto q \wedge M$ and we apply the inequality above to $q_M \in L^\infty$. Taking the limit $M \to \infty$ yields the desired result. □

8.4.2 Dissipation of energy

Proposition 8.29. *The generalized free energy decreases along the generalized MFS dynamics $(p_t)_{t\geqslant 0}$ solution to (8.9). More precisely, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{F}^{\sigma,\gamma}(p_t) = -\int_{\mathbb{R}^d} \left|\frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t,x)\right|^2 p_t(x)\,\mathrm{d}x. \tag{8.36}$$

Proof. Using Theorem 8.28 whose assumptions are satisfied in view of Theorem 8.7 and Lemma 8.24, we have

$$\begin{aligned}\mathfrak{F}^{\sigma,\gamma}(p_{t+h}) - \mathfrak{F}^{\sigma,\gamma}(p_t) &\geqslant \int_{\mathbb{R}^d} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t,x)(p_{t+h}-p_t)(x)\,\mathrm{d}x \\ &= -\int_{\mathbb{R}^d} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t,x)\int_t^{t+h} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_s,x)p_s(x)\,\mathrm{d}s\,\mathrm{d}x.\end{aligned}$$

Similarly we have

$$\mathfrak{F}^{\sigma,\gamma}(p_{t+h}) - \mathfrak{F}^{\sigma,\gamma}(p_t) \leqslant -\int_{\mathbb{R}^d} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t+h},x)\int_t^{t+h} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_s,x)p_s(x)\,\mathrm{d}s\,\mathrm{d}x.$$

The conclusion then follows from the dominated convergence theorem. Indeed, by Theorem 8.7, the mapping $t \mapsto \delta\mathfrak{F}^{\sigma,\gamma}/\delta p\,(p_t,x)$ is continuous and satisfies

$$\sup_{t\leqslant T}\left|\frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t,x)\right| \leqslant C_T(1+|x|^2)$$

for any $T > 0$. Note that the same holds for $\delta F/\delta p\,(p_t, x)$ by the $\mathcal{W}_1$-continuity of $t \mapsto p_t$. In addition, Lemma 8.24 ensures that $\int |x|^4 \sup_{t\leqslant T} p_t(x)\,\mathrm{d}x < \infty$. $\square$

The dissipation of energy allows us to extend previous estimates of the generalized MFS dynamics from $[0,T]$ to $[0,\infty)$ which is crucial to study its asymptotic behavior.

Lemma 8.30. *It holds*

$$\sup_{t>0}\left\{\int_{\mathbb{R}^d} |x|^2 p_t(x)\,\mathrm{d}x + \int_{\mathbb{R}^d} |\nabla\sqrt{p_t}(x)|^2\,\mathrm{d}x\right\} < +\infty. \tag{8.37}$$

Proof. Let q the Gaussian density with variance υ^2. We have

$$H(p) = H(p\,|\,q) + \int p(x)\log q(x)\,\mathrm{d}x \geqslant -\frac{d}{2}\log(2\pi\upsilon^2) - \frac{1}{2\upsilon^2}\int_{\mathbb{R}^d}|x|^2 p(x)\,\mathrm{d}x.$$

Then it follows from Assumption 8.3 by choosing υ sufficiently large that there exist C, $c > 0$ such that

$$\mathfrak{F}^{\sigma,\gamma}(p_t) \geqslant -C + c\int_{\mathbb{R}^d}|x|^2 p_t(x)\,\mathrm{d}x + \sigma^2\int_{\mathbb{R}^d}|\nabla\sqrt{p_t}(x)|^2\,\mathrm{d}x, \qquad \forall t \geqslant 0. \tag{8.38}$$

Since the generalized free energy is decreasing according to Theorem 8.29, we deduce that

$$\sup_{t\geqslant 0}\left\{c\int_{\mathbb{R}^d}|x|^2 p_t(x)\,\mathrm{d}x + \sigma^2\int_{\mathbb{R}^d}|\nabla\sqrt{p_t}(x)|^2\,\mathrm{d}x\right\} \leqslant C + \mathfrak{F}^{\sigma,\gamma}(p_0). \qquad \square$$

Proposition 8.31. *It holds for all $x \in \mathbb{R}^d$,*

$$\sup_{t\geqslant 0}|\nabla\log p_t(x)| \leqslant C(1+|x|).$$

Proof. In view of Theorem 8.7, the Hessian $\nabla^2\log p_t$ is bounded by some constant, denoted L. In particular, it holds

$$|\nabla\log p_t(x)| \leqslant L|x| + |\nabla\log p_t(0)|,$$

and also

$$|\nabla\log p_t(0)|^2 \leqslant \big(L|x| + |\nabla\log p_t(x)|\big)^2 \leqslant 2L^2|x|^2 + 2|\nabla\log p_t(x)|^2.$$

It follows that

$$\begin{aligned}4\int_{\mathbb{R}^d}|\nabla\sqrt{p_t}(x)|^2\,\mathrm{d}x &= \int_{\mathbb{R}^d}|\nabla\log p_t(x)|^2 p_t(x)\,\mathrm{d}x\\ &\geqslant \frac{1}{2}|\nabla\log p_t(0)|^2 - L^2\int_{\mathbb{R}^d}|x|^2 p_t(x)\,\mathrm{d}x.\end{aligned}$$

We conclude by Lemma 8.30 that $\sup_{t\geqslant 0}|\nabla\log p_t(0)| < \infty$. $\square$

Using Lemma 8.31, it is straightforward to extend the Gaussian bounds of Lemma 8.24 from $[0,T]$ to $\mathbb{R}_+$.

Corollary 8.32. *There exist $\underline{c}$, $\bar{c}$, $\underline{C}$, $\bar{C} > 0$ such that for all $t \geqslant 0$, $x \in \mathbb{R}^d$,*

$$\underline{C}\exp(-\underline{c}|x|^2) \leqslant p_t(x) \leqslant \bar{C}\exp(-\bar{c}|x|^2).$$

8.4.3 Proof of Theorem 8.8

Proof of Theorem 8.8. We start by observing that the family $(p_t)_{t\geqslant 0}$ is relatively compact for the uniform norm on $C(\mathbb{R}^d)$. This property follows from Arzelà–Ascoli Theorem as

$$p_t(x) \leqslant Ce^{-c|x|^2}, \qquad |\nabla p_t(x)| = |\nabla \log p_t(x)| p_t(x) \leqslant C(1+|x|)\exp(-c|x|^2), \tag{8.39}$$

by Lemma 8.31 and Lemma 8.32. Let p^* be an arbitrary cluster point, i.e., p_{t_k} converges uniformly to p^* for some sequence $t_k \uparrow \infty$. Note that, in view of the Gaussian bound above, the convergence also occurs in $\mathcal{W}_p$ for any $p \geqslant 1$. The aim of the proof is to show that p^* is the unique minimizer of $\mathfrak{F}^{\sigma,\gamma}$.

Step 1. Let us show first that, for almost all $h > 0$,

$$\liminf_{k\to\infty} \int_{\mathbb{R}^d} \left| \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t_k+h}, x) \right|^2 p(t_k+h, x)\,\mathrm{d}x = 0. \tag{8.40}$$

Indeed, suppose by contradiction that there exists $h > 0$ such that

$$\begin{aligned} 0 &< \int_0^h \liminf_{k\to\infty} \left\{ \int_{\mathbb{R}^d} \left| \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t_k+s}, x) \right|^2 p_{t_k+s}(x)\,\mathrm{d}x \right\} \mathrm{d}s \\ &\leqslant \liminf_{k\to\infty} \int_0^h \left\{ \int_{\mathbb{R}^d} \left| \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t_k+s}, x) \right|^2 p_{t_k+s}(x)\,\mathrm{d}x \right\} \mathrm{d}s, \end{aligned}$$

where the last inequality is due to Fatou's lemma. It would lead to a contradiction as by Theorem 8.29,

$$\begin{aligned} \mathfrak{F}^{\sigma,\gamma}(p_{t_{k+1}}) - \mathfrak{F}^{\sigma,\gamma}(p_{t_0}) &= \sum_{j=0}^{k} \mathfrak{F}^{\sigma,\gamma}(p_{t_{j+1}}) - \mathfrak{F}^{\sigma,\gamma}(p_{t_j}) \\ &= -\sum_{j=0}^{k} \int_0^{t_{j+1}-t_j} \int_{\mathbb{R}^d} \left| \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t_j+s}, x) \right|^2 p_{t_j+s}(x)\,\mathrm{d}x\,\mathrm{d}s \end{aligned}$$

where the l.h.s. is bounded from below by (8.38) and the r.h.s. diverges to $-\infty$ by assuming w.l.o.g. that $t_{j+1} - t_j \geqslant h$.

Step 2. From now on, denote by $t_k^h := t_k + h$ where $h > 0$ is chosen so that (8.40) holds. Let q be an arbitrary probability measure in $\mathcal{P}_H$. Due to the first order inequality established in Theorem 8.28, we have

$$\mathfrak{F}^{\sigma,\gamma}(q) - \mathfrak{F}^{\sigma,\gamma}(p_{t_k^h}) \geqslant \int_{\mathbb{R}^d} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t_k^h}, x)(q - p_{t_k^h})(x)\,\mathrm{d}x.$$

In view of Theorem 8.7, Lemma 8.31 and Lemma 8.32, we have

$$\sup_{t\geqslant 0} \left| \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_t, x) \right| \leqslant C(1+|x|^2), \qquad \sup_{t\geqslant 0} \int_{\mathbb{R}^d} |x|^2 p_t(x)\,\mathrm{d}x < \infty.$$

Note that the first inequality holds for $\delta F/\delta p\,(p_t, x)$ since $(p_t)_{t\geqslant 0}$ belongs to a $\mathcal{W}^1$-compact set due to the Gaussian bound. Hence, for any $\varepsilon > 0$, we can find K big enough such that for all k, $j \in \mathbb{N}$,

$$\mathfrak{F}^{\sigma,\gamma}(p_{t_k^h}) \leqslant \mathfrak{F}^{\sigma,\gamma}(q) - \int_{|x|\leqslant K} \frac{\delta \mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t_k^h}, x)(q - p_{t_k^h})(x)\,\mathrm{d}x + \varepsilon.$$

Further it follows from Cauchy–Schwartz inequality that

$$\Big|\int_{|x|\leqslant K}\frac{\delta\mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t^h_k},x)(q-p_{t^h_k})(x)\,\mathrm{d}x\Big|$$
$$\leqslant\Big(\int_{\mathbb{R}^d}\Big|\frac{\delta\mathfrak{F}^{\sigma,\gamma}}{\delta p}(p_{t^h_k},x)\Big|^2 p_{t^h_k}(x)\,\mathrm{d}x\int_{|x|\leqslant K}\frac{|q-p_{t^h_k}|^2}{p_{t^h_k}}(x)\,\mathrm{d}x\Big)^{1/2}$$

Assume first that q is bounded and note that the second term on the r.h.s. is also bounded as $\inf_{k,h,x}p_{t^h_k}(x)>0$ by Lemma 8.32. Thus we deduce by taking the limit $k\to\infty$ and then $\varepsilon\to 0$ that

$$\liminf_{k\to\infty}\mathfrak{F}^{\sigma,\gamma}(p_{t^h_k})\leqslant\mathfrak{F}^{\sigma,\gamma}(q),\tag{8.41}$$

for any $q\in\mathcal{P}_H$ bounded. If $q\in\mathcal{P}_H$ is not necessarily bounded, this inequality also holds as it holds for the distribution $q_M\propto q\wedge M$ and $\mathfrak{F}^{\sigma,\gamma}(q_M)\to\mathfrak{F}^{\sigma,\gamma}(q)$ as $M\to\infty$.

Step 3. Denote by $(p^*_t)_{t\geqslant 0}$ the solution to (8.9) starting from $p^*_0=p^*$ We observe by Lemma 8.33 below that $p_{t^h_k}$ and $\nabla\log p_{t^h_k}$ converges pointwise to p^*_h and $\nabla\log p^*_h$ respectively. In view of Lemma 8.31 and Lemma 8.32, it follows easily by the dominated convergence theorem that $\lim_{k\to\infty}F(p_{t^h_k})=F(p^*_h)$ as $p_{t^h_k}\to p^*_h$ in $\mathcal{W}_2$ by using the Gaussian bound,

$$\lim_{k\to\infty}H(p_{t^h_k})=\lim_{k\to\infty}\int p_{t^h_k}\log p_{t^h_k}=\int p^*_h\log p^*_h=H(p^*_h),$$

and

$$\lim_{k\to\infty}I(p_{t^h_k})=\lim_{k\to\infty}\frac14\int|\nabla\log p_{t^h_k}|^2p_{t^h_k}=\frac14\int|\nabla\log p^*_h|^2p^*_h=I(p^*_h).$$

We deduce that

$$\lim_{k\to\infty}\mathfrak{F}^{\sigma,\gamma}(p_{t^h_k})=\mathfrak{F}^{\sigma,\gamma}(p^*_h).$$

Hence, by (8.41), p^*_h is a minimizer of $\mathfrak{F}^{\sigma,\gamma}$. In view of Theorem 8.26, this minimizer is unique and thus p^*_h does not depend on h and coincides with its limit $p^*_0=p^*$ when $h\to 0$.

Step 4. As a byproduct, we observe that p^* is a stationary solution to (8.9) and thus it satisfies

$$\frac{\delta\mathfrak{F}^{\sigma,\gamma}}{\delta p}(p^*,\cdot)=0.\qquad\square$$

Lemma 8.33. *Using the notations above, as $k\to\infty$, $p_{t^h_k}$ converges uniformly to p^*_h and $\nabla\log p_{t^h_k}$ converges to $\nabla\log p^*_h$ in $\|\cdot\|_{(2)}$.*

Proof. Step 1. Let us show first that $\nabla\log p_{t_k}$ converges to $\nabla\log p^*$ in $\|\cdot\|_{(2)}$. According to Theorem 8.7 and Lemma 8.31, $(\nabla\log p_{t_k})_{k\in\mathbb{N}}$ lives in a $\|\cdot\|_{(2)}$-compact set of the form

$$\mathcal{K}\coloneqq\{f:\mathbb{R}^d\to\mathbb{R}\mid f\text{ is }C\text{-Lipschitz and }|f(0)|\leqslant C\},$$

for some constant $C > 0$. Consequently, there is a subsequence and a function $f \in \mathcal{K}$ such that $\lim_{k\to\infty} \|\nabla \log p_{t_k} - f\|_{(2)} = 0$. Therefore, we have for almost all $x,\, y \in \mathbb{R}^d$,

$$\begin{aligned}\log p^*(x) - \log p^*(y) &= \lim_{k\to\infty} \big(\log p_{t_k}(x) - \log p_{t_k}(y)\big)\\ &= \lim_{k\to\infty} \int_0^1 \nabla \log p_{t_k}\big(sx + (1-s)y\big) \cdot (x-y)\,\mathrm{d}s\\ &= \int_0^1 f\big(sx + (1-s)y\big) \cdot (x-y)\,\mathrm{d}s.\end{aligned}$$

So $f = \nabla \log p^*$ and the desired result follows.

Step 2. In view of Proposition 8.25, it follows immediately from Step (i) that $(p_{t_k^h}, \nabla \log p_{t_k^h})$ converges to $(p_h^*, \nabla \log p_h^*)$ in $\mathcal{W}_1 \otimes \|\cdot\|_{(2)}$. It remains to prove that $p_{t_k^h}$ converges uniformly to p_h^*. This is an easy consequence of Arzelà–Ascoli Theorem by (8.39). $\square$

8.4.4 Proof of Theorem 8.11

The proof relies on the following functional inequality which is new to the best of our knowledge and may carry independent interest.

Theorem 8.34. *Let $p(\mathrm{d}x) = e^{-u(x)}\,\mathrm{d}x$ satisfy a Poincaré inequality with constant C_P, i.e., for all $f \in H^1(p)$ such that $\int f\,\mathrm{d}p = 0$,*

$$\int f^2\,\mathrm{d}p \leqslant C_P \int |\nabla f|^2\,\mathrm{d}p. \tag{8.42}$$

Assume that u is weakly differentiable with $\nabla u \in L^2$ and define the operator $\mathcal{L} := \Delta - \nabla u \cdot \nabla$. Then we have for all $f \in W^{2,2}(p)$ such that $\mathcal{L}f \in L^2(p)$,

$$\begin{aligned}&C_P^{-1} \left(\int_{\mathbb{R}^d} f(x)p(\mathrm{d}x)\right)^2 \int_{\mathbb{R}^d} |\nabla f(x)|^2 p(\mathrm{d}x)\\ &\qquad\leqslant \int_{\mathbb{R}^d} f(x)^2 p(\mathrm{d}x) \int_{\mathbb{R}^d} \big(\mathcal{L}f(x)\big)^2 p(\mathrm{d}x) - \left(\int_{\mathbb{R}^d} f(x)\mathcal{L}f(x)p(\mathrm{d}x)\right)^2.\end{aligned} \tag{8.43}$$

Remark 8.35. Note that it follows from integration by parts that

$$\int_{\mathbb{R}^d} \mathcal{L}f(x)p(\mathrm{d}x) = 0, \qquad \int_{\mathbb{R}^d} |\nabla f(x)|^2 p(\mathrm{d}x) = -\int_{\mathbb{R}^d} f(x)\mathcal{L}f(x)p(\mathrm{d}x). \tag{8.44}$$

Moreover, if $p^f(\mathrm{d}x) = f(x)^2 p(\mathrm{d}x)$ is a probability measure then the right hand side of the inequality (8.43) is equal to the variance of $\mathcal{L}f/f$ under p^f, namely, $\mathrm{Var}_{p^f}(\mathcal{L}f/f)$.

Proof of Theorem 8.34. Let $f = f_0 + \bar{f}$, where $\bar{f} = \int f\,\mathrm{d}p$ is the mean. For the right-hand side of the inequality (8.43), we obtain by using successively $\int f_0\,\mathrm{d}p = 0$, $\int \mathcal{L}f\,\mathrm{d}p = 0$ and Cauchy–Schwartz inequality,

$$\begin{aligned}&\int_{\mathbb{R}^d} f^2\,\mathrm{d}p \int_{\mathbb{R}^d} (\mathcal{L}f)^2\,\mathrm{d}p - \left(\int_{\mathbb{R}^d} f\mathcal{L}f\,\mathrm{d}p\right)^2\\ &\qquad= \bar{f}^2 \int (\mathcal{L}f)^2\,\mathrm{d}p + \int f_0^2\,\mathrm{d}p \int (\mathcal{L}f)^2\,\mathrm{d}p - \left(\int f_0\mathcal{L}f\,\mathrm{d}p\right)^2 \geqslant \bar{f}^2 \int (\mathcal{L}f)^2\,\mathrm{d}p.\end{aligned}$$

Meanwhile for the left-hand side, we obtain by (8.44), Cauchy–Schwarz inequality and Poincaré inequality,

$$\begin{aligned}\int |\nabla f|^2 \,\mathrm{d}p &= -\int f\mathcal{L}f \,\mathrm{d}p = -\int f_0\mathcal{L}f \,\mathrm{d}p \\ &\leqslant \Big(\int f_0^2 \,\mathrm{d}p\Big)^{1/2}\Big(\int (\mathcal{L}f)^2 \,\mathrm{d}p\Big)^{1/2} \leqslant C_{\mathrm{P}}^{1/2}\Big(\int |\nabla f|^2 \,\mathrm{d}p \int (\mathcal{L}f)^2 \,\mathrm{d}p\Big)^{1/2}.\end{aligned}$$

The desired inequality follows by combining the estimates above. □

Proposition 8.36. *If $u : \mathbb{R}^d \to \mathbb{R}$ decomposes as $u = v + w$ with $v, w \in \mathcal{C}^2$, $\nabla^2 v \geqslant \eta I_d$ with $\eta > 0$ and $|\nabla w| \leqslant L$, then there exists a constant $C_P = C(\eta, L, d)$ such that the Poincaré inequality (8.42) holds.*

Proof. This is a direct consequence of Corollary 1.6 (1) in [9]. □

Proof of Theorem 8.11. Recall that p_t is the classical solution to the MFS dynamics (8.10). For each $t > 0$, denote $F_t \coloneqq \delta F/\delta p\,(p_t, \cdot)$ and define

$$\hat{p}_t \coloneqq \operatorname*{argmin}_{p\in\mathcal{P}_H}\Big\{\int F_t \,\mathrm{d}p + \frac{\sigma^2}{4}I(p)\Big\}. \tag{8.45}$$

We recognize that it is the minimizer of the mean field optimization problem if we replace $F(p)$ by $\int F_t \,\mathrm{d}p$. According to Theorem 8.8, the minimizer $\hat{p}_t = e^{-\hat{u}_t}$ satisfies $\hat{u}_t = \hat{v}_t + \hat{w}_t$ with $\nabla^2 \hat{v}_t \geqslant \underline{\eta} I_d$ and $|\nabla \hat{w}_t| \leqslant L$ for all $t > 0$. Thus $\hat{p}_t$ verifies a Poincaré inequality with a constant C_P independent of time by Proposition 8.36. Note also that

$$\frac{\sigma^2}{2}\Delta \hat{u}_t - \frac{\sigma^2}{4}|\nabla \hat{u}_t|^2 + F_t - \hat{\lambda}_t = 0, \tag{8.46}$$

where, by integration by parts,

$$\hat{\lambda}_t = \int \Big(\frac{\sigma^2}{2}\Delta \hat{u}_t - \frac{\sigma^2}{4}|\nabla \hat{u}_t|^2 + F_t\Big) \,\mathrm{d}\hat{p}_t = \int \Big(\frac{\sigma^2}{4}|\nabla \hat{u}_t|^2 + F_t\Big) \,\mathrm{d}\hat{p}_t. \tag{8.47}$$

The desired result follows by applying the functional inequality (8.43) with distribution $\hat{p}_t$ and function $f_t = \sqrt{p_t/\hat{p}_t}$. Let $\mathcal{L}_t = \Delta - \nabla \hat{u}_t \cdot \nabla$ and observe by direct computation using $f_t = \exp\big((\hat{u}_t - u_t)/2\big)$ that

$$\frac{\mathcal{L}_t f_t}{f_t} = \frac{1}{2}\Delta \hat{u}_t - \frac{1}{4}|\nabla \hat{u}_t|^2 - \Big(\frac{1}{2}\Delta u_t - \frac{1}{4}|\nabla u_t|^2\Big).$$

Then it follows from (8.46) that

$$\frac{\mathcal{L}_t f_t}{f_t} = \sigma^{-2}\hat{\lambda}_t - \sigma^{-2}\Big(\frac{\sigma^2}{2}\Delta u_t - \frac{\sigma^2}{4}|\nabla u_t|^2 + F_t\Big). \tag{8.48}$$

Thus, by using Theorem 8.29, the right-hand side of (8.43) corresponds to

$$\begin{aligned}\frac{\mathrm{d}\mathfrak{F}^\sigma(p_t)}{\mathrm{d}t} &= -\int \Big|\frac{\sigma^2}{2}\Delta u_t - \frac{\sigma^2}{4}|\nabla u_t|^2 + F_t - \lambda_t\Big|^2 \,\mathrm{d}p_t \\ &= -\int \Big|\frac{\sigma^2}{2}\Delta u_t - \frac{\sigma^2}{4}|\nabla u_t|^2 + F_t - \hat{\lambda}_t\Big|^2 \,\mathrm{d}p_t + (\hat{\lambda}_t - \lambda_t)^2 \\ &= -\sigma^4 \operatorname{Var}_{p_t}\Big(\frac{\mathcal{L}_t f_t}{f_t}\Big),\end{aligned}$$

where, by integration by parts,

$$\lambda_t = \int \left(\frac{\sigma^2}{2} \Delta u_t - \frac{\sigma^2}{4} |\nabla u_t|^2 + F_t \right) \mathrm{d}p_t = \int \left(\frac{\sigma^2}{4} |\nabla u_t|^2 + F_t \right) \mathrm{d}p_t. \tag{8.49}$$

As for the left-hand side of (8.43), we have for the first term

$$\int f_t \, \mathrm{d}\hat{p}_t = \int \sqrt{p_t \hat{p}_t} \, \mathrm{d}x \geqslant C > 0,$$

by using the Gaussian bounds provided in Lemma 8.32. Regarding the second term, it holds by using (8.44) and (8.48),

$$\int |\nabla f_t|^2 \, \mathrm{d}\hat{p}_t = -\int f_t \mathcal{L}_t f_t \, \mathrm{d}\hat{p}_t = \sigma^{-2} \int \left(\frac{\sigma^2}{2} \Delta u_t - \frac{\sigma^2}{4} |\nabla u_t|^2 + F_t \right) \mathrm{d}p_t - \sigma^{-2} \hat{\lambda}_t.$$

Using further (8.47) and (8.49), we obtain

$$\begin{aligned} \sigma^2 \int |\nabla f_t|^2 \, \mathrm{d}\hat{p}_t &= \int \left(\frac{\sigma^2}{4} |\nabla u_t|^2 + F_t \right) \mathrm{d}p_t - \int \left(\frac{\sigma^2}{4} |\nabla \hat{u}_t|^2 + F_t \right) \mathrm{d}\hat{p}_t \\ &= \int F_t (\mathrm{d}p_t - \mathrm{d}\hat{p}_t) + \frac{\sigma^2}{4} \big(I(p_t) - I(\hat{p}_t) \big) \\ &\geqslant \int F_t (\mathrm{d}p_t - \mathrm{d}p^*) + \frac{\sigma^2}{4} \big(I(p_t) - I(p^*) \big), \end{aligned}$$

where the last inequality follows from the optimality of $\hat{p}_t$ in (8.45).

By Theorem 8.34 and the above computations, we deduce that

$$\begin{aligned} \frac{\mathrm{d}\mathfrak{F}^\sigma(p_t)}{\mathrm{d}t} &\leqslant -\frac{(C\sigma)^2}{C_P} \left(\int F_t (\mathrm{d}p_t - \mathrm{d}p^*) + \frac{\sigma^2}{4} \big(I(p_t) - I(p^*) \big) \right) \\ &\leqslant -\frac{(C\sigma)^2}{C_P} \big(\mathfrak{F}^\sigma(p_t) - \mathfrak{F}^\sigma(p^*) \big), \end{aligned}$$

where the last inequality is due to Theorem 8.28. Therefore, the exponential convergence of the free energy (8.13) follows with a constant $c = (C\sigma)^2 / C_P$.

In order to obtain the exponential convergence of the relative Fisher information, define $f_t^* \coloneqq \sqrt{p_t / p^*}$, $\mathcal{L}^* \coloneqq \Delta - \nabla u^* \cdot \nabla$, and repeat the previous computation:

$$\begin{aligned} I(p_t | p^*) &= 4 \int |\nabla f_t^*|^2 \, \mathrm{d}p^* = -4 \int f_t^* \mathcal{L}^* f_t^* \, \mathrm{d}p^* \\ &= 4\sigma^{-2} \left(\int \frac{\delta F}{\delta p}(p^*, \cdot)(\mathrm{d}p_t - \mathrm{d}p^*) + \frac{\sigma^2}{4} (I(p_t) - I(p^*)) \right) \\ &\leqslant 4\sigma^{-2} \big(\mathfrak{F}^\sigma(p_t) - \mathfrak{F}^\sigma(p^*) \big). \end{aligned}$$

□

8.5 Gradient flow with relative entropy

Let p_i^h be defined in (8.15). The proof of Theorem 8.13 essentially relies on applying Arzelà–Ascoli Theorem to the family $(t, x) \mapsto p^h_{\lfloor t/h \rfloor}(x)$ for $h > 0$. To this end, we need to ensure equicontinuity and boundedness in the two subsequent sections. In the sequel, we fix a time horizon $T < \infty$ and we denote by $N_h \coloneqq \lfloor T/h \rfloor$.

8.5.1 Equicontinuity in space

The goal of this section is to obtain uniform Gaussian bounds for the family $(p_i^h)_{h,i\leqslant N_h}$ as in Lemma 8.24 and to deduce equicontinuity in space of the discrete flow.

Proposition 8.37. *For some $\underline{C}$, $\bar{C}$, $\underline{c}$, $\bar{c} > 0$, we have for all $h > 0$, $i \leqslant N_h$, $x \in \mathbb{R}^d$,*

$$\underline{C}\exp(-\underline{c}|x|^2) \leqslant p_i^h(x) \leqslant \bar{C}\exp(-\bar{c}|x|^2).$$

In addition, it holds

$$\sup_{h,i\leqslant N_h} \|\nabla p_i^h\|_\infty < +\infty.$$

Proof. The Gaussian bounds are a direct consequence of Lemma E.4, whose assumptions are satisfied according to Lemmas 8.38–8.42 below. As for the second part, it follows from the identity $\nabla p_i^h = p_i^h \nabla \log p_i^h$ by using the Gaussian upper-bound above and the fact that $|\nabla \log p_i^h(x)| \leqslant C(1+|x|)$ according to Lemmas 8.41 and 8.42 below. □

Recall that the mapping p_i^h is a solution to the stationary MFS equation (8.16). In other words, if we denote $u_i^h := -\log(p_i^h)$, it holds

$$\frac{\sigma^2}{2}\Delta u_i^h - \frac{\sigma^2}{4}|\nabla u_i^h|^2 + \frac{\delta F}{\delta p}(p_i^h,\cdot) + h^{-1}u_{i-1}^h - h^{-1}u_i^h = \lambda_i^h, \tag{8.50}$$

with

$$\lambda_i^h = \int_{\mathbb{R}^d}\left(\frac{\delta F}{\delta p}(p_i^h,\cdot) + h^{-1}(u_{i-1}^h - u_i^h) + \frac{\sigma^2}{2}\Delta u_i^h - \frac{\sigma^2}{4}|\nabla u_i^h|^2\right)p_i^h. \tag{8.51}$$

The key point is to observe that we have the decomposition $u_i^h = v_i^h + w_i^h$ with v_i^h uniformly convex and w_i^h uniformly Lipschitz. It comes from using arguments similar to Section 8.3.1. In this setting there is a slight ambiguity in the definition of v_i^h (and thus w_i^h) due to the normalizing constant λ_i^h. Let us define v_i^h as the solution to

$$\frac{\sigma^2}{2}\Delta v_i^h - \frac{\sigma^2}{4}|\nabla v_i^h|^2 + g + h^{-1}v_{i-1}^h - h^{-1}v_i^h = 0.$$

Lemma 8.38. *The function $(v_i^h)_{h,i\leqslant N_h}$ are uniformly η-convex for some $\eta > 0$.*

Proof. Observe that v_i^h corresponds to the stationary solution to (8.21) with parameter $\gamma = h^{-1}$ and convex term $g + h^{-1}v_{i-1}^h$ instead of g. Due to Proposition 8.20, v_i^h is η_i^h-convex with

$$\begin{aligned}
\eta_i^h &= \frac{\sqrt{h^{-2} + 4\sigma^2\big(\underline{\kappa} + h^{-1}\eta_{i-1}^h\big)} - h^{-1}}{2\sigma^2}\\
&\geqslant \frac{\sqrt{h^{-2} + 4\sigma^2\big(\underline{\kappa} + h^{-1}\min(\eta_{i-1}^h, \sqrt{\underline{\kappa}}/\sigma)\big)} - h^{-1}}{2\sigma^2}\\
&\geqslant \min(\eta_{i-1}^h, \sqrt{\underline{\kappa}}/\sigma).
\end{aligned}$$

Recall that $\eta_0^h = \underline{\eta}_0$. Finally we obtain that v_i^h is $\min(\underline{\eta}_0, \sqrt{\underline{\kappa}}/\sigma)$-convex. □

Lemma 8.39. *The Hessian's $(\nabla^2 v_i^h)_{h,i\leqslant N_h}$ are uniformly bounded.*

Proof. As in Proposition 8.20, we may obtain the following probabilistic representation:

$$\nabla v_i^h(x) = \mathbb{E}\Big[\int_0^t e^{-s/h}\Big(\nabla g(X_s) + h^{-1}\nabla v_{i-1}^h(X_s)\Big)\,\mathrm{d}s + e^{-t/h}\nabla v_i^h(X_t)\Big],$$

with

$$X_s = x - \int_0^s \frac{\sigma^2}{2}\nabla v_i^h(X_r)\,\mathrm{d}r + \sigma W_s.$$

Let X' satisfy the same SDE with initial value x'. Since v_i^h is η-convex, it follows from the same arguments as (8.30) that

$$|X_t - X_t'| \leqslant e^{-\sigma^2\eta t/2}|x - x'|.$$

Further we obtain

$$\begin{aligned}
&|\nabla v_i^h(x) - \nabla v_i^h(x')| \\
&\leqslant \mathbb{E}\Big[\int_0^t e^{-s/h}(\bar\kappa + h^{-1}\|\nabla^2 v_{i-1}^h\|_\infty)|X_s - X_s'|\,\mathrm{d}s + e^{-t/h}\|\nabla^2 v_i^h\|_\infty|X_t - X_t'|\Big] \\
&\leqslant \Big(\int_0^t e^{-(1/h+\sigma^2\eta/2)s}(\bar\kappa + h^{-1}\|\nabla^2 v_{i-1}^h\|_\infty)\,\mathrm{d}s + e^{-(1/h+\sigma^2\eta/2)t}\|\nabla^2 v_i^h\|_\infty\Big)|x - x'|.
\end{aligned}$$

Letting $t \to \infty$, we get

$$\|\nabla^2 v_i^h\|_\infty \leqslant \frac{\bar\kappa h + \|\nabla^2 v_{i-1}^h\|_\infty}{1 + \sigma^2\eta h/2}.$$

Therefore, we deduce by induction that

$$\|\nabla^2 v_i^h\|_\infty \leqslant \frac{2\bar\kappa}{\sigma^2\eta}\Big(1 - \Big(1 + \frac{\sigma^2\eta h}{2}\Big)^{-i}\Big) + \bar\eta_0\Big(1 + \frac{\sigma^2\eta h}{2}\Big)^{-i} \leqslant \frac{2\bar\kappa}{\sigma^2\eta} + \bar\eta_0. \qquad \square$$

Lemma 8.40. *The gradients $(\nabla w_i^h)_{h,i\leqslant N_h}$ are uniformly bounded.*

Proof. Observe that $w_i^h = u_i^h - v_i^h$ satisfies

$$\frac{\sigma^2}{2}\Delta w_i^h - \frac{\sigma^2}{2}\nabla v_i^h\cdot\nabla w_i^h - \frac{\sigma^2}{4}|\nabla w_i^h|^2 + G(p_i^h,\cdot) + h^{-1}w_{i-1}^h - h^{-1}w_i^h = \lambda_i^h.$$

As in Proposition 8.21, we observe that w_i^h is the value function of the following stochastic control problem

$$\begin{aligned}
w_i^h(x) = \inf_\alpha \mathbb{E}\Big[\int_0^t e^{-s/h}\Big(G(p_i^h, X_s^\alpha) + h^{-1}w_{i-1}^h(X_s^\alpha) + \frac{\sigma^2}{4}|\alpha_s|^2 - \lambda_i^h\Big)\,\mathrm{d}s \\
+ e^{-t/h}w_i^h(X_t^\alpha)\Big],
\end{aligned}$$

with

$$\mathrm{d}X_s^\alpha = -\frac{\sigma^2}{2}\big(\nabla v_i^h(X_s^\alpha) + \alpha_s\big)\,\mathrm{d}s + \sigma\,\mathrm{d}W_s, \quad X_0^\alpha = x.$$

Further as in (8.31), we may estimate

$$
\begin{aligned}
&|w_i^h(x) - w_i^h(x')| \\
&\leqslant \left(\int_0^t e^{-(1/h+\sigma^2\eta/2)s}(L_G + h^{-1}\|\nabla w_{i-1}^h\|_\infty) + e^{-(1/h+\sigma^2\eta/2)t}\|\nabla w_i^h\|_\infty \right)|x - x'|.
\end{aligned}
$$

Letting $T \to \infty$, we obtain

$$
\|\nabla w_i^h\|_\infty \leqslant \frac{L_G h + \|\nabla w_{i-1}^h\|_\infty}{1 + \sigma^2\eta h/2}.
$$

Therefore, we deduce by induction that

$$
\|\nabla w_i^h\|_\infty \leqslant \frac{2L_G}{\sigma^2\eta} + L_0. \qquad \square
$$

Lemma 8.41. *The Hessians* $(\nabla^2 u_i^h)_{h,i\leqslant N_h}$ *are uniformly bounded.*

Proof. As in the proof of Lemma 8.22, the Feynman–Kac formula ensures that

$$
\nabla u_i^h(x) = \mathbb{E}\Big[\int_0^\infty e^{-t/h}\Big(\nabla\frac{\delta F}{\delta p}(p_i^h, X_t) + h^{-1}\nabla u_{i-1}^h(X_t)\Big)\,\mathrm{d}t\Big],
$$

with

$$
X_t = x - \frac{\sigma^2}{2}\int_0^t \nabla u_i^h(X_s)\,\mathrm{d}s + \sigma W_t.
$$

Let Y satisfy the same SDE starting from y. By the reflection coupling in Theorem E.7, it holds

$$
\mathcal{W}_1(p_t^X, p_t^Y) \leqslant Ce^{-ct}|x - y|,
$$

where p^X and p^Y are the marginal distribution of X and Y respectively. Then it follows by Kantorovich duality that

$$
\begin{aligned}
&|\nabla u_i^h(x) - \nabla u_i^h(y)| \\
&\quad \leqslant \int_0^\infty Ce^{-t/h-ct}|x-y| + \int_0^\infty e^{-t/h}h^{-1}\mathbb{E}\big[|\nabla u_{i-1}^h(X_t) - \nabla u_{i-1}^h(Y_t)|\big] \\
&\quad = \frac{Ch}{1+ch}|x-y| + \int_0^\infty e^{-t/h}h^{-1}\mathbb{E}\big[|\nabla u_{i-1}^h(X_t) - \nabla u_{i-1}^h(Y_t)|\big]\,\mathrm{d}t.
\end{aligned}
$$

Next apply the same estimate on $|\nabla u_{i-1}^h(X_t) - \nabla u_{i-1}^h(Y_t)|$, and obtain

$$
\begin{aligned}
&|\nabla u_i^h(x) - \nabla u_i^h(y)| \leqslant \frac{2Ch}{1+ch}|x-y| \\
&\quad + \int_0^\infty e^{-t_1/h}h^{-1}\int_0^\infty e^{-t_2/h}h^{-1}\mathbb{E}\big[|\nabla u_{i-2}^h(X_{t_1+t_2}^{(1)}) - \nabla u_{i-2}^h(Y_{t_1+t_2}^{(1)})|\big]\,\mathrm{d}t_2\,\mathrm{d}t_1,
\end{aligned}
$$

with

$$
X_0^{(1)} = x, \quad \mathrm{d}X_t^{(1)} = \begin{cases} -\frac{\sigma^2}{2}\nabla u_i^h(X_t^{(1)})\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, & \text{for } t \in [0, t_1) \\ -\frac{\sigma^2}{2}\nabla u_{i-1}^h(X_t^{(1)})\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, & \text{for } t \geqslant t_1. \end{cases}
$$

By repeating the procedure, we eventually obtain for $i \geqslant 1$

$$
\begin{aligned}
&|\nabla u_i^h(x) - \nabla u_i^h(y)| \\
&\quad \leqslant \frac{Chi}{1+ch}|x-y| \\
&\quad + \int_0^\infty \cdots \int_0^\infty e^{-h^{-1}\sum_{j=1}^i t_j} h^{-i} \mathbb{E}\Big[\big|\nabla u_0\big(X^{(i-1)}_{\sum_{j=1}^i t_j}\big) - \nabla u_0\big(Y^{(i-1)}_{\sum_{j=1}^i t_j}\big)\big|\Big] \,\mathrm{d}t_i \cdots \mathrm{d}t_1,
\end{aligned}
$$

with

$$
X_0^{(i-1)} = x, \qquad \mathrm{d}X_t^{(i-1)} = -\frac{\sigma^2}{2}\nabla u_j^h\big(X_t^{(i-1)}\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad \text{for } t \in [t_{i-j}, t_{i+1-j}).
$$

Again it follows from the reflection coupling that

$$
\mathcal{W}_1(p_t^{X^{(i-1)}}, p_t^{Y^{(i-1)}}) \leqslant Ce^{-ct}|x-y|,
$$

where $p^{X^{(i-1)}}$, $p^{Y^{(i-1)}}$ are the marginal distribution of $X^{(i-1)}$, $Y^{(i-1)}$ respectively. In particular, the constants c, C do not depend on $(t_1, \cdots, t_{i-1})$ by Lemmas 8.38–8.40. Finally we get

$$
\begin{aligned}
&|\nabla u_i^h(x) - \nabla u_i^h(y)| \\
&\quad \leqslant \frac{Chi}{1+ch}|x-y| + C\int_0^\infty \cdots \int_0^\infty e^{-(h^{-1}+c)\sum_{j=1}^i t_j} h^{-i}|x-y| \,\mathrm{d}t_i \cdots \mathrm{d}t_1 \\
&\quad \leqslant C(T+1)|x-y|,
\end{aligned}
$$

and the desired result follows. □

Lemma 8.42. *The vectors $\big(\nabla u_i^h(0)\big)_{h, i \leqslant N_h}$ are uniformly bounded.*

Proof. The proof follows similar arguments as Lemma 8.30 and Lemma 8.31. First we observe that the sequence $\mathfrak{F}^\sigma(p_i^h)$ is non-increasing as

$$
\mathfrak{F}^\sigma(p_i^h) \leqslant \mathfrak{F}^\sigma(p_i^h) + h^{-1}H(p_i^h|p_{i-1}^h) \leqslant \mathfrak{F}^\sigma(p_{i-1}^h) + h^{-1}H(p_{i-1}^h|p_{i-1}^h) = \mathfrak{F}^\sigma(p_{i-1}^h),
$$

by using (8.15) for the second inequality. In addition, it follows from Assumption 8.3 that

$$
\lambda\int_{\mathbb{R}^d}|x|^2 p_i^h(x)\,\mathrm{d}x + \sigma^2\int_{\mathbb{R}^d}|\nabla\sqrt{p_i^h}(x)|^2\,\mathrm{d}x \leqslant \mathfrak{F}^\sigma(p_i^h).
$$

Therefore we have

$$
\sup_{h, i\leqslant N_h}\left\{\lambda\int_{\mathbb{R}^d}|x|^2 p_i^h(x)\,\mathrm{d}x + \sigma^2\int_{\mathbb{R}^d}|\nabla\sqrt{p_i^h}(x)|^2\,\mathrm{d}x\right\} \leqslant \mathfrak{F}^\sigma(p_0).
$$

Since we have proved that $L := \sup_{h,i\leqslant N_h}\|\nabla^2 u_i^h\|_\infty < \infty$, we deduce that

$$
4\int_{\mathbb{R}^d}|\nabla\sqrt{p_i^h}(x)|^2\,\mathrm{d}x = \int_{\mathbb{R}^d}|\nabla u_i^h(x)|^2 p_i^h(x)dx \geqslant \frac{1}{2}|\nabla u_i^h(0)|^2 - L^2\int_{\mathbb{R}^d}|x|^2 p_i^h(x)dx.
$$

Finally we obtain $\sup_{h,i\leqslant N_h}|\nabla u_i^h(0)| < \infty$. □

8.5.2 Equicontinuity in time

We aim to show the equicontinuity in time of the family $(p^h)_{h>0}$ as stated in the proposition below. We also demonstrate as a preliminary step and for later use that the family of function $(t \mapsto \lambda^h_{\lfloor t/h \rfloor})_{h>0}$ defined by (8.51) is bounded and equicontinuous.

Proposition 8.43. *There exists constants $C, c > 0$ such that for all $h > 0$, $i < j \leqslant N_h$, $x \in \mathbb{R}^d$,*

$$|p^h_j(x) - p^h_i(x)| \leqslant C \exp(-c|x|^2)(j-i)h.$$

Additionally, the sequence $(\lambda^h_i)_{h.i \leqslant N_h}$ is uniformly bounded, i.e., $\sup_{h,i\leqslant N_h} |\lambda^h_i| < +\infty$, and there exists a modulus of continuity (m.o.c.) $\varpi : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $h > 0$, $i < j \leqslant N_h$,

$$|\lambda^h_j - \lambda^h_i| \leqslant \varpi\big((j-i)h\big).$$

Proof. Step 1: Formulas for λ^h_i. The normalization condition for u^h_i, $i \leqslant N_h$, writes

$$\begin{aligned} 1 = \int \exp(-u^h_i) &= \int \exp(-u^h_{i-1}) \exp\left(-h \frac{u^h_i - u^h_{i-1}}{h}\right) \\ &= \int p^h_{i-1} \exp\left(-h\left(\frac{\sigma^2}{2}\Delta u^h_i - \frac{\sigma^2}{4}|\nabla u^h_i|^2 + \frac{\delta F}{\delta p}(p^h_i, \cdot) - \lambda^h_i\right)\right). \end{aligned}$$

where the latter follows from (8.50). This allows us to obtain the following formula for λ^h_i:

$$\lambda^h_i = -\frac{1}{h} \log \int p^h_{i-1} \exp(-hB^h_i), \tag{8.52}$$

where

$$B^h_i := \frac{\sigma^2}{2}\Delta u^h_i - \frac{\sigma^2}{4}|\nabla u^h_i|^2 + \frac{\delta F}{\delta p}(p^h_i, \cdot).$$

By writing the normalization in the backward way,

$$\begin{aligned} 1 = \int \exp(-u^h_{i-1}) = \int \exp(-u^h_i) \exp\left(h \frac{u^h_i - u^h_{i-1}}{h}\right) \\ = \int \exp(-u^h_i) \exp\big(h(B^h_i - \lambda^h_i)\big), \end{aligned}$$

we obtain a similar formula

$$\lambda^h_i = \frac{1}{h} \log \int p^h_i \exp(hB^h_i). \tag{8.53}$$

We apply Jensen's inequality to (8.52) and (8.53) to obtain

$$\int p^h_i B^h_i \leqslant \lambda^h_i \leqslant \int p^h_{i-1} B^h_i. \tag{8.54}$$

Additionally, estimates from Lemma 8.41 and Lemma 8.42 gives us the bound

$$\sup_{h,i\leqslant N_h} |B^h_i(x)| \leqslant C(1 + |x|^2). \tag{8.55}$$

Note that the same holds for $\delta F/\delta p\,(p_i^h,\cdot)$ as $(p_i^h)_{h,i\leqslant N_h}$ belong to a $\mathcal{W}_1$-compact set due to the Gaussian bound. Thus, by Corollary 8.37, we prove the second claim $\sup_{h,i\leqslant N_h}|\lambda_i^h|<+\infty$.

Step 2: Time regularity of p_i^h. According to the HJB equation (8.50) and Step (i) above, it holds

$$|u_j^h(x)-u_i^h(x)|=\left|h\Big(\sum_{s=i+1}^{j}B_s^h-\sum_{s=i+1}^{j}\lambda_s^h\Big)\right|\leqslant C(j-i)h(1+|x|^2). \tag{8.56}$$

Using further the bound from Corollary 8.37, we obtain

$$\begin{aligned}|p_j^h(x)-p_i^h(x)|&=\big|\exp\big(-u_j^h(x)\big)-\exp\big(-u_i^h(x)\big)\big|\\&\leqslant p_j^h(x)\vee p_i^h(x)\,|u_j^h(x)-u_i^h(x)|\\&\leqslant C(j-i)h\exp(-c|x|^2)(1+|x|^2)\\&\leqslant C(j-i)h\exp(-c|x|^2),\end{aligned}\tag{8.57}$$

which is our first claim. This implies the $\mathcal{W}_1$-regularity of p_i^h as follows:

$$\mathcal{W}_1(p_j^h,p_i^h)\leqslant\int|x||p_j^h(x)-p_i^h(x)|\,\mathrm{d}x\leqslant C(j-i)h\int|x|\exp(-c|x|^2)\leqslant C(j-i)h. \tag{8.58}$$

Step 3: Uniform continuity of $\delta F/\delta p$. Thanks to the estimate in Corollary 8.37, $\{p_i^h\}_{h,i\leqslant N_h}$ forms a relatively compact set in $\mathcal{W}_1$, and the $\mathcal{W}_1$-continuity of $p\mapsto\delta F/\delta p\,(p,0)$ becomes uniform. That is, there exists a m.o.c. $\varpi_0:\mathbb{R}_+\to\mathbb{R}_+$ such that

$$\left|\frac{\delta F}{\delta p}(p_i^h,0)-\frac{\delta F}{\delta p}(p_j^h,0)\right|\leqslant\varpi_0\big(\mathcal{W}_1(p_i^h,p_j^h)\big),\qquad\forall h>0,\ \forall i\leqslant N_h,\ \forall j\leqslant N_h.$$

Integrating along the straight line from 0 to any $x\in\mathbb{R}^d$ and using the assumptions on $\nabla\frac{\delta F}{\delta p}$, we obtain

$$\begin{aligned}\left|\frac{\delta F}{\delta p}(p_i^h,x)-\frac{\delta F}{\delta p}(p_j^h,x)\right|&\leqslant\left|\frac{\delta F}{\delta p}(p_i^h,0)-\frac{\delta F}{\delta p}(p_j^h,0)\right|\\&\qquad+\int_0^1\left|x\cdot\left(\nabla\frac{\delta F}{\delta p}(p_i^h,tx)-\nabla\frac{\delta F}{\delta p}(p_j^h,tx)\right)\right|\mathrm{d}t\\&\leqslant\varpi_0\big(\mathcal{W}_1(p_i^h,p_j^h)\big)+L_G|x|\mathcal{W}_1(p_i^h,p_j^h).\end{aligned}$$

Combining with (8.58), we deduce that the exists a m.o.c. $\varpi_1:\mathbb{R}_+\to\mathbb{R}_+$ such that

$$\left|\frac{\delta F}{\delta p}(p_i^h,x)-\frac{\delta F}{\delta p}(p_j^h,x)\right|\leqslant(1+|x|)\varpi_1\big((j-i)h\big). \tag{8.59}$$

Step 4: Time regularity of λ_i^h. We first note that thanks to (8.54) we can approximate λ_i^h by $\int p_i^hB_i^h$, up to a uniform $O(h)$ error. More precisely,

$$|r_i^h|:=\left|\lambda_i^h-\int p_i^hB_i^h\right|\leqslant\left|\int(p_i^h-p_{i-1}^h)B_i^h\right|\leqslant Ch\int\exp(-c|x|^2)(1+|x|^2)\leqslant Ch,$$

where we used (8.55) and (8.57). It suffices then to study the difference

$$\int p_j^h B_j^h - p_i^h B_i^h = \int p_i^h (B_j^h - B_i^h) + \int (p_j^h - p_i^h) B_j^h =: \delta + \delta'.$$

We bound the second part, using again (8.55) and (8.57),

$$|\delta'| \leqslant C(j-i)h \int \exp(-c|x|^2)(1+|x|^2) \leqslant C(j-i)h.$$

As for the first part, we decompose it into three terms, each of which we treat separately:

$$\delta = \frac{\sigma^2}{2} \int p_i^h (\Delta u_j^h - \Delta u_i^h) - \frac{\sigma^2}{4} \int p_i^h (|\nabla u_j^h|^2 - |\nabla u_i^h|^2) \\ + \int p_i^h \left(\frac{\delta F}{\delta p}(p_j^h, \cdot) - \frac{\delta F}{\delta p}(p_i^h, \cdot) \right) =: \delta_1 + \delta_2 + \delta_3.$$

We apply integration by parts to the first term, using the previous estimates on ∇u_i^h, p_i^h and the time regularity result of ∇u_i^h from Lemma 8.44 below,

$$|\delta_1| = \frac{\sigma^2}{2} \left| \int p_i^h \nabla u_i^h \cdot (\nabla u_j^h - \nabla u_i^h) \right| \leqslant C \int p_i^h (1+|x|)^2 \big((j-i)h\big)^{1/2} \leqslant C\big((j-i)h\big)^{1/2}.$$

The second term is treated in the same way:

$$|\delta_2| \leqslant \frac{\sigma^2}{4} \int p_i^h (|\nabla u_j^h| + |\nabla u_i^h|) |\nabla u_j^h - \nabla u_i^h| \leqslant C\big((j-i)h\big)^{1/2}.$$

Using (8.59), we can then bound

$$|\delta_3| \leqslant \int p_i^h \left| \frac{\delta F}{\delta p}(p_j^h, \cdot) - \frac{\delta F}{\delta p}(p_i^h, \cdot) \right| \leqslant \int p_i^h (1+|x|) \varpi_1\big((j-i)h\big) \leqslant C \varpi_1\big((j-i)h\big).$$

Collecting the bounds on r, δ', δ, we derive finally that

$$|\lambda_j^h - \lambda_i^h| \leqslant |\delta| + |\delta'| + |r_j^h| + |r_i^h| \\ \leqslant C\Big(\big(2(j-i)h\big)^{1/2} + \varpi_1\big((j-i)h\big) + (j-i)h + 2h\Big). \qquad \square$$

Lemma 8.44. *There exists a constant C such that for all $h \in (0,1)$, $i < j \leqslant N_h$, we have*

$$|\nabla u_j^h(x) - \nabla u_i^h(x)| \leqslant C\big((j-i)h\big)^{1/2}(1+|x|), \qquad \forall x \in \mathbb{R}^d.$$

Proof. By taking spatial derivatives of the HJB equation (8.50), we see the following is satisfied for

$$\frac{1}{h}(\nabla u_k^h - \nabla u_{k-1}^h) = \frac{\sigma^2}{2} \Delta \nabla u_k^h - \frac{\sigma^2}{2} \nabla^2 u_k^h \nabla u_k^h + \nabla \frac{\delta F}{\delta p}(p_k^h, \cdot) =: \frac{\sigma^2}{2} \Delta \nabla u_k^h + A_k^h, \tag{8.60}$$

where by estimates in Lemma 8.41 and Lemma 8.42 we know that

$$\sup_{h, i \leqslant N_h} |A_i^h(x)| \leqslant C(1+|x|), \quad \forall x \in \mathbb{R}^d.$$

The solution to (8.60) admits the following representation

$$\nabla u_k^h = \int_0^\infty e^{-h^{-1}t}\Big(P_{\sigma^2 t}A_k^h + \frac{1}{h}P_{\sigma^2 t}\nabla u_{k-1}^h\Big)\,\mathrm{d}t,$$

where P_t is the heat kernel generated by $\frac{1}{2}\Delta$. Iterating this procedure with descending k, we obtain

$$\begin{aligned}\nabla u_j^h = \sum_{n=1}^{j-i} h^{-(n-1)}\int_{t_1,\dots,t_n\geqslant 0} & e^{-h^{-1}(t_1+\dots+t_n)}P_{\sigma^2(t_1+\dots+t_n)}A_{j+1-n}^h\,\mathrm{d}t_1\cdots\mathrm{d}t_n\\ + h^{-(j-i)}\int_{t_1,\dots,t_{j-i}\geqslant 0} & e^{-h^{-1}(t_1+\dots+t_{j-i})}P_{\sigma^2(t_1+\dots+t_{j-i})}\nabla u_i^h\,\mathrm{d}t_1\cdots\mathrm{d}t_{j-i}.\end{aligned}$$

Here we used the semigroup property of the heat kernel. Denoting $\gamma_{n,\theta}(t) = \Gamma(n)^{-1}\theta^{-n}t^{n-1}e^{-t/\theta}$ the gamma distribution density, we have equivalently

$$\nabla u_j^h = h\sum_{n=1}^{j-i}\int_0^\infty \gamma_{n,h}(t)P_{\sigma^2 t}A_{j+1-n}^h\,\mathrm{d}t + \int_0^\infty \gamma_{j-i,h}(t)P_{\sigma^2 t}\nabla u_i^h\,\mathrm{d}t.$$

Subtracting ∇u_i^h, we obtain

$$\begin{aligned}&|\nabla u_j^h(x) - \nabla u_i^h(x)|\\ &\leqslant h\sum_{n=1}^{j-i}\int_0^\infty \gamma_{n,h}(t)|P_{\sigma^2 t}A_{j+1-n}^h(x)|\,\mathrm{d}t + \int_0^\infty \gamma_{j-i,h}(t)|P_{\sigma^2 t}\nabla u_i^h(x) - \nabla u_i^h(x)|\,\mathrm{d}t\\ &\leqslant h\sum_{n=1}^{j-i}\int_0^\infty \gamma_{n,h}(t)C\big(1+|x|+(\sigma^2 t)^{1/2}\big)\,\mathrm{d}t + \int_0^\infty \gamma_{j-i,h}(t)(\sigma^2 t)^{1/2}\|\nabla^2 u_i^h\|_\infty\,\mathrm{d}t\\ &\leqslant C(j-i)h(1+|x|) + Ch^{3/2}\sum_{n=1}^{j-i}\frac{\Gamma\big(n+\frac{1}{2}\big)}{\Gamma(n)} + Ch^{1/2}\frac{\Gamma\big(j-i+\frac{1}{2}\big)}{\Gamma(j-i)}\\ &\leqslant C(j-i)h(1+|x|) + C\big((j-i)h\big)^{3/2} + C\big((j-i)h\big)^{1/2}.\end{aligned}$$

In the second inequality, we used the following properties of the heat kernel:

$$P_t|\cdot|(x) \leqslant c_d\sqrt{t} + |x|, \quad \|P_t f - f\|_\infty \leqslant \sqrt{t}\|f\|_{\mathrm{Lip}}.$$

In the last inequality, we used the log-convexity of the gamma function along the positive real line: $\Gamma\big(x+\frac{1}{2}\big) \leqslant \sqrt{\Gamma(x)\Gamma(x+1)} = \sqrt{x}\Gamma(x)$ for $x > 0$. $\square$

8.5.3 Proof of Theorem 8.13

Proof of Theorem 8.13. Step 1. Let us define by abuse of notations the step flows

$$f^h(t) = f_i^h, \qquad \text{for } t \in [ih,(i+1)h), \qquad f = p, \lambda.$$

In view of Corollary 8.37 and Lemma 8.43, we can apply a version of Arzelà–Ascoli Theorem for discontinuous functions, see e.g. [74, Theorem 6.1], to ensure that the family of functions $(p^h)_h$ (resp. $(\lambda^h)_h$) is relatively compact in $B([0,T]\times\mathbb{R}^d)$ (resp. $B([0,T])$) the space of bounded functions on $[0,T]\times\mathbb{R}^d$ (resp. $[0,T]$) equipped with the uniform norm, and any adherence values p (resp. λ) is uniformly continuous.

Let p and λ be such adherence values, i.e., there exists $h_n \downarrow 0$ such that $p^{h_n} \to p$ and $\lambda^{h_n} \to \lambda$ uniformly. Note that $\psi^{h_n} := \sqrt{p^{h_n}}$ also converges to $\psi := \sqrt{p}$ uniformly on $[0,T] \times \mathbb{R}^d$ by using the elementary inequality $|\sqrt{a} - \sqrt{b}| \leqslant \sqrt{|a-b|}$.

Step 2. Let us verify that the limit (p, ψ, λ) solves the MFS equation (8.17) in the weak sense, i.e., for all $\varphi \in C_c^2(\mathbb{R}^d)$, we have for all $t \in [0,T]$,

$$\begin{aligned}\int \big(\psi(t,x) - \psi(0,x)\big)\varphi(x)\,\mathrm{d}x \\ = \int_0^t \int \frac{\sigma^2}{2}\psi(s,x)\Delta\varphi(x) - \frac{1}{2}\bigg(\frac{\delta F}{\delta p}(p_s,x) - \lambda(s)\bigg)\psi(s,x)\varphi(x)\,\mathrm{d}x\,\mathrm{d}s. \end{aligned} \tag{8.61}$$

By construction, we know that the following holds for $i \leqslant N_h$,

$$\begin{aligned}\int \sum_{k=1}^{i} \log \frac{\psi^h(kh,x)}{\psi^h\big((k-1)h,x\big)}\psi^h(kh,x)\varphi(x)\,\mathrm{d}x \\ = h\sum_{k=1}^{i}\int \frac{\sigma^2}{2}\psi^h(kh,x)\Delta\varphi(x) - \frac{1}{2}\bigg(\frac{\delta F}{\delta p}(p^h_{kh},x) - \lambda^h(kh)\bigg)\psi^h(kh,x)\varphi(x)\,\mathrm{d}x. \end{aligned} \tag{8.62}$$

Let $i = \lfloor t/h \rfloor$ be the unique integer such that $t \in [ih, (i+1)h)$ and denote the difference between the left and right hand sides of (8.61), (8.62) by $\delta^\ell(h)$, $\delta^r(h)$ respectively. We want to show that both $\delta^\ell(h_n)$, $\delta^r(h_n)$ converge to zero when $n \to \infty$, so that (8.61) is proved. For the left hand side we have $\delta^\ell(h) = \delta^\ell_1(h) + \delta^\ell_2(h)$ with

$$\begin{aligned}\delta^\ell_1(h) &= \int \big(\psi(t,x) - \psi^h(t,x)\big)\varphi(x)\,\mathrm{d}x, \\ \delta^\ell_2(h) &= \int \sum_{k=1}^{i} \bigg(\psi^h(kh,x) - \psi^h\big((k-1)h,x\big) \\ &\qquad - \log\frac{\psi^h(kh,x)}{\psi^h((k-1)h,x)}\psi^h(kh,x)\bigg)\varphi(x)\,\mathrm{d}x.\end{aligned}$$

The first part converges to 0 along the sequence h_n as $\psi^{h_n} \to \psi$ uniformly. For the second part we note that, by using (8.56),

$$\begin{aligned}&\bigg|\psi^h(kh,x) - \psi^h\big((k-1)h,x\big) - \log\frac{\psi^h(kh,x)}{\psi^h\big((k-1)h,x\big)}\psi^h(kh,x)\bigg| \\ &= \bigg|\exp\big(-u^h_k(x)/2\big) - \exp\big(-u^h_{k-1}(x)/2\big) + \frac{1}{2}\exp\big(-u^h_k(x)/2\big)\big(u^h_k(x) - u^h_{k-1}(x)\big)\bigg| \\ &\leqslant \frac{1}{8}\max\big(\psi^h_k(x), \psi^h_{k-1}(x)\big)|u^h_k(x) - u^h_{k-1}(x)|^2 \leqslant C\exp(-c|x|^2)h^2,\end{aligned}$$

so that $\delta^\ell_2(h) \leqslant Ch\int \exp(-c|x|^2)\varphi(x)\,\mathrm{d}x \leqslant Ch$. For the right hand side, we have

$\delta^r(h) = \delta_1^r(h) + \delta_2^r(h)$ with

$$\delta_1^r(h) = \int_{ih}^{t} \int \frac{\sigma^2}{2} \psi(s,x) \Delta\varphi(x) - \frac{1}{2}\Big(\frac{\delta F}{\delta p}(p_s, x) - \lambda(s)\Big)\psi(s,x)\varphi(x)\, \mathrm{d}x\, \mathrm{d}s,$$

$$\begin{aligned}\delta_2^r(h) = & \int_0^{ih}\int \frac{\sigma^2}{2}(\psi - \psi^h)(s,x)\Delta\varphi(x) \\ & - \frac{1}{2}\left(\frac{\delta F}{\delta p}(p_\cdot, \cdot)\psi - \frac{\delta F}{\delta p}(p_\cdot^h, \cdot)\psi^h - \lambda\psi + \lambda^h\psi^h\right)(s,x)\varphi(x)\, \mathrm{d}x\, \mathrm{d}s,\end{aligned}$$

The first part clearly satisfies $|\delta_1^r(h)| \leqslant Ch$ while the second part goes to zero along the sequence h_n as $(p^{h_n}, \psi^{h_n}, \lambda^{h_n}) \to (p, \psi, \lambda)$ uniformly.

Step 3. If we denote $c(t,x) \coloneqq \delta F/\delta p\,(p_t, x) - \lambda(t)$, then Step 2 ensures that ψ is a weak solution to the linear PDE

$$\partial_t \psi_t = \frac{\sigma^2}{2}\Delta\psi_t - \frac{1}{2}c_t\psi_t.$$

By weak uniqueness and strong existence, it is actually the classical solution to this PDE. It follows that $p_t = \psi_t^2$ satisfies (8.10) with $\lambda_t = \lambda(p_t)$ as the mass of p_t is conserved to 1 by construction. We conclude by uniqueness stated in Theorem 8.7. $\square$

Appendix A

Appendices to Chapter 1

A.1 Proofs of technical results on MFL

In the section we provide proofs of technical results on the regularity properties of the MFL dynamics.

Proof of Proposition 1.37. It is classical that under the conditions (1.3) and (1.5), the McKean–Vlasov SDE

$$\mathrm{d}X_t = -D_m F(m_t, X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \qquad \operatorname{Law}(X_t) = m_t$$

has unique global solution defined for $t \in [0, +\infty)$. By construction the marginal law $m_t = \operatorname{Law}(X_t)$ is in $C\big([0,+\infty); \mathcal{P}_2(\mathbb{R}^d)\big)$, proving the existence of solution. Any solution to the Fokker–Planck equation admits equally this probabilistic representation, then the uniqueness in short time follows from Cauchy–Lipschitz bounds. We extend this uniqueness to the infinity by sewing up the short time intervals, finishing the proof of the first claim.

Let $\rho_t(x)$ be the density of Gaussian $\mathcal{N}(0, 2t)$. The solution m_t satisfies Duhamel's formula in the sense of distributions

$$\begin{aligned} m_t &= \rho_t \star m_0 + \int_0^t \rho_{t-s} \star \nabla \cdot \big(m_s D_m F(m_s, \cdot)\big)\,\mathrm{d}s \\ &= \rho_t \star m_0 + \sum_{i=1}^d \int_0^t \nabla_i \rho_{t-s} \star \big(m_s D_m F^i(m_s, \cdot)\big)\,\mathrm{d}s. \end{aligned}$$

Note that $\|\nabla \rho_t\|_{L^p(\mathbb{R}^d)} \leqslant C_{d,p} t^{-\frac{1}{2}+\frac{d}{2}(\frac{1}{p}-1)}$, which is integrable around $0+$ when $p < \frac{d}{d-1}$. In this case apply Young's convolution inequality to obtain

$$\|m_t\|_{L^p(\mathbb{R}^d)} \leqslant \|\rho_t\|_{L^p(\mathbb{R}^d)} \|m_0\|_{\mathrm{TV}} + \sum_{i=1}^d \int_0^t \|\nabla_i \rho_{t-s}\|_{L^p(\mathbb{R}^d)} \|m_s D_m F^i(m_s, \cdot)\|_{\mathrm{TV}}\,\mathrm{d}s,$$

where $\sup_{s\in[0,t]} \|m_s D_m F^i(m_s,\cdot)\|_{\mathrm{TV}} \leqslant \sup_{s\in[0,t]} C \int (1+|x|) m_s(\mathrm{d}x) < +\infty$. Hence $\|m_t\|_{L^p(\mathbb{R}^d)} < +\infty$ for all $t > 0$. This and the second moment bound $\int |x|^2 m_t(\mathrm{d}x) < +\infty$ are sufficient for the finiteness of entropy, i.e. the integral $\int |\log m_t(x)| m_t(x)\,\mathrm{d}x$ is finite, which is our second claim. Indeed for the lower bound on entropy we use the decomposition in (1.45), while the upper bounds follows from $m\log m \leqslant \frac{m^p - m}{p-1}$ for all $p > 1$.

The drift $D_m F(m_t, x)$ has uniform linear growth in x:

$$|D_m F(m_s, x)| \leqslant M^F_{mx}|x| + \sup_{s\in[t_0,t]} |D_m F(m_s, 0)|,$$

where M^F_{mx} is the constant in (1.5) and the second term is finite by the compactness of set $\{m_s : s \in [t_0, t]\}$ in $\mathcal{P}_2$. As a result,

$$\int_{t_0}^t \int |D_m F(m_s, x)|^2 m_s(\mathrm{d}x)\,\mathrm{d}t < +\infty.$$

We then apply [22, Theorem 7.4.1] to obtain the finiteness of (1.47). Especially, $\nabla m \in L^1_{\mathrm{loc}}\big((0, +\infty); L^1(\mathbb{R}^d)\big)$. Rewrite the Fokker–Planck equations as a continuity equation $\partial_t m + \nabla\cdot(m_t v_t) = 0$ where $v_t(x) = -D_m F(m_t, x) - \nabla \log m_t(x)$. We have

$$\begin{aligned}
&\int_{t_0}^t \int |v_s(x)|^2 m_s(\mathrm{d}x)\,\mathrm{d}s \\
&\qquad \leqslant 2\bigg(\int_{t_0}^t \int |D_m F(m_s, x)|^2 m_s(\mathrm{d}x)\,\mathrm{d}s + \int_{t_0}^t \int \frac{|\nabla m_s(x)|^2}{m_s(x)}\,\mathrm{d}x\,\mathrm{d}s\bigg) < +\infty.
\end{aligned}$$

Hence by [4, Theorem 8.3.1] the flow m_t is locally AC^2 in $(\mathcal{P}_2, W_2)$. The vector field $v_t(x) = -D_m F(m_t, x) - \nabla \log m_t(x)$ solves the continuity equation

$$\partial_t m_t + \nabla\cdot(m_t v_t) = 0 \tag{A.1}$$

in the sense of distributions and v_t writes in the gradient form $v_t = -\nabla\big(\frac{\delta F}{\delta m}(m_t, x) + \log m_t(x)\big) = -\nabla\varphi_t$.

We finally verify v_t is indeed a tangent vector of m_t according to [4, Definition 8.4.1], i.e. $v_t \in \mathrm{Tan}_{m_t}\mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\varphi : \varphi \in C^\infty_c(\mathbb{R}^d)\}}^{L^2(m_t)}$. Let $\eta_R : \mathbb{R}^d \to [0, 1]$ be a smooth function supported on $B(2R)$, has the constant value 1 on $B(R)$ and satisfies $|\nabla\eta(x)| \leqslant 2/R$ for all x. We have

$$\int |\nabla\varphi_t - \nabla(\varphi_t\eta_R)|^2 m_t \leqslant 2\int_{B(2R)\setminus B(R)} \big(|\varphi_t|^2|\nabla\eta_R|^2 + |\nabla\varphi_t|^2|1-\eta_R|^2\big)m_t.$$

The second term tends to 0 when $R \to \infty$, while the first satisfies

$$\begin{aligned}
&\int_{B(2R)\setminus B(R)} |\varphi_t|^2|\nabla\eta_R|^2 m_t \\
&\quad \leqslant \frac{2}{R^2}\int_{B(2R)\setminus B(R)} \bigg(\bigg|\frac{\delta F}{\delta m}(m_t, x)\bigg|^2 + |\log m_t(x)|^2\bigg)m_t \\
&\quad \leqslant \frac{2C}{R^2}\int_{B(2R)\setminus B(R)} (1 + |x|^4)m_t(\mathrm{d}x) + \frac{2}{R^2}\int_{B(2R)\setminus B(R)} |\log m_t|^2 m_t \\
&\quad \leqslant \frac{2C}{R^2}\int_{B(2R)\setminus B(R)} (1 + 4R^2|x|^2)m_t(\mathrm{d}x) + \frac{2}{R^2}\int_{B(2R)\setminus B(R)} |\log m_t|^2 m_t.
\end{aligned}$$

Here the first term tends to 0 since $m_t \in \mathcal{P}_2$, while the second term tends to 0 by the integrability of $|\log m_t|^2 m_t$, which follows from the elementary inequality

$$m|\log m|^2 \leqslant C_p m^p \mathbb{1}_{m\geqslant 1} + 2\Big(|x|^2 m + \sup_{t\in[0,1]} t(\log t)^2 e^{-|x|}\Big)\mathbb{1}_{m<1}$$

for $p > 1$ and $x \in \mathbb{R}^d$. Hence $\nabla(\varphi_t \eta_R) \to \nabla\varphi_t$ in $L^2(m_t)$. It then suffices to approximate the (essentially) compactly supported function $\varphi_t\eta_R$ by C_c^∞ functions in the $L^2(m_t)$-norm. We can do this by taking a sequence of compacted supported mollifiers ρ_n and applying them to obtain $\nabla(\varphi_t\eta_R)\star\rho_n \to \nabla(\varphi_t\eta_R)$ in $L^2(m_t)$ when $n \to \infty$. $\square$

Proof of Proposition 1.43. Let h be a positive function. Define the functions $k_n = \mathbb{1}_{B(n)}(h \wedge n) \vee 1/n$ and $k_{n,m} = \rho_m \star k_n$, where $(\rho_m)_{m\in\mathbb{N}}$ is a sequence of C^∞ mollifiers. They satisfy

$$\forall x \in \mathbb{R}^d, \qquad \frac{1}{n} \leqslant k_n(x), k_{n,m}(x) \leqslant n \text{ and } |\nabla^\ell k_{n,m}(x)| \leqslant n\|\nabla^\ell \rho_m\|_\infty < +\infty.$$

In particular $k_{n,m} \in \mathcal{A}_+$. We have $k_n \to h$ in $L^p(\mu)$ whenever $h \in L^p(\mu)$ for $p \geqslant 1$ and $\|k_n\|_q \to \|h\|_q$ whenever $h \in L^q(\mu)$ for $q \leqslant 1$ by the dominated convergence theorem. Since for all $n \in \mathbb{N}$ the function $k_n \in L^1(\mathbb{R}^d)$, we have $k_{n,m} \to k_n$ in $L^1(\mathbb{R}^d)$ when $m \to \infty$. Hence $k_{n,m} \to k_n$ a.e. when $m \to \infty$ along a subsequence. Then we can apply again the dominated convergence to obtain $k_{n,m} \to k_n$ in $L^p(\mu)$ for all $p \geqslant 1$ and $\|k_{n,m}\|_q \to \|k_n\|_q$ for all $q < 1$. We can thus taking a subsequence of $(n,m) \to (+\infty,+\infty)$ so that $k_{n,m} \to h$ in the desired ways. $\square$

Proof of Proposition 1.44. Fix $T > t_0$. We denote by C a positive constant that depends on $\max_{k=1,2,3}\sup_{m,x}|\nabla^k D_m F(m,x)|$ and on the initial condition $h' \in \mathcal{A}_+$; and by C_Q a positive constant that depends additionally on the quantity Q. The constants C, C_Q may change from line to line. Define $g(t,x) = \nabla\cdot(b_t - b_\infty) + (b_t - b_\infty)\cdot b_\infty$. It satisfies $|g(t,x)| \leqslant C(1+|x|)$ for all $(t,x) \in [t_0,T]\times\mathbb{R}^d$ as $\|\nabla^k(b_t - b_\infty)\|_\infty \leqslant C$ for $k = 0, 1$ and $t \in [t_0,T]$. Fix $t \in [t_0,T]$. Let $(X_s^{t,x})_{s\in[0,t-t_0]}$ be the stochastic process solving

$$\mathrm{d}X_s^{t,x} = (2b_\infty - b_{t-s})\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W_s \tag{A.2}$$

with $X_0^{t,x} = x$ and define as well its extremal process $M_s^{t,x} = \sup_{0\leqslant u\leqslant s}|X_u|$ for $s \in [0,t-t_0]$. Since the drift satisfies $(2b_\infty - b_t)\cdot x \leqslant C_T|x|^2 + C_T$ for all $(t,x) \in [t_0,T]\times\mathbb{R}^d$, we obtain the Gaussian moment bound

$$\mathbb{E}\exp\bigl(C_T^{-1}|M_{t-t_0}^{t,x}|^2\bigr) \leqslant C_T\exp(C_T|x|^2)$$

by Itō's formula and Doob's maximal inequality. As a consequence the exponential moments are finite:

$$\forall\alpha \geqslant 0, \qquad \mathbb{E}\exp\bigl(\alpha|M_{t-t_0}^{t,x}|\bigr) \leqslant C_{T,\alpha}\exp(C_{T,\alpha}|x|).$$

Set $h(t_0,\cdot) = h'$. We construct the solution by the Feynman–Kac formula for (1.50)

$$h(t,x) \coloneqq \mathbb{E}\biggl[\exp\biggl(-\int_0^{t-t_0} g(t-s,X_s^{t,x})\,\mathrm{d}s\biggr)h(t_0,X_{t-t_0}^{t,x})\biggr].$$

It is standard that the h constructed above solves (1.50) in the sense of distributions. We verify $h_t \in \mathcal{A}_+$ for all $t \in [t_0,T]$. For the upper bound we apply the Cauchy–Schwarz inequality to obtain

$$\begin{aligned} h(t,x) &\leqslant \mathbb{E}\biggl[\exp\biggl(-2\int_0^{t-t_0} g(t-s,X_s^{t,x})\,\mathrm{d}s\biggr)\biggr]^{1/2}\mathbb{E}\bigl[h(t_0,X_{t-t_0}^{t,x})^2\bigr]^{1/2} \\ &\leqslant \mathbb{E}\bigl[\exp\bigl(C_T(1+|M_{t-t_0}^{t,x}|)\bigr)\bigr]^{1/2}\,\mathbb{E}\bigl[\exp\bigl(C_T(1+|X_{t-t_0}^{t,x}|)\bigr)\bigr]^{1/2} \\ &\leqslant \mathbb{E}\bigl[\exp\bigl(C_T(1+|M_{t-t_0}^{t,x}|)\bigr)\bigr] \leqslant \exp\bigl(C_T(1+|x|)\bigr). \end{aligned}$$

We applied the bound on g and h in the second inequality and used the exponential moment bound on M_{t-t_0} in the last. For the lower bound we use Cauchy–Schwarz from the other direction:

$$\begin{aligned} h(t,x) &\geqslant \mathbb{E}\biggl[\exp\biggl(\int_0^{t-t_0} g(t-s,X_s^{t,x})\,\mathrm{d}s\biggr)\biggr]^{-1}\mathbb{E}\bigl[h(t_0,X_{t-t_0}^{t,x})^{1/2}\bigr]^2 \\ &\geqslant C_T^{-1}\,\mathbb{E}\bigl[\exp\bigl(C_T|M_{t-t_0}^{t,x}|\bigr)\bigr]^{-1}\,\mathbb{E}\bigl[\exp\bigl(-C_T|X_{t-t_0}^{t,x}|\bigr)\bigr]^2 \\ &\geqslant C_T^{-1}\,\mathbb{E}\bigl[\exp\bigl(C_T|M_{t-t_0}^{t,x}|\bigr)\bigr]^{-1}\,\mathbb{E}\bigl[\exp\bigl(C_T|X_{t-t_0}^{t,x}|\bigr)\bigr]^{-2} \\ &\geqslant C_T^{-1}\,\mathbb{E}\bigl[\exp\bigl(C_T|M_{t-t_0}^{t,x}|\bigr)\bigr]^{-3} \geqslant C_T^{-1}\exp(-C_T|x|). \end{aligned}$$

Again we applied the bound on g and h on the second inequality and used the exponential moment bound on M_{t-t_0} on the last line. So we have proved the bound of both sides $|\log h(t,x)| \leqslant C_T(1+|x|)$, that is, the "zeroth-order" condition of $\mathcal{A}_+$.

Now derive the continuity of $x \mapsto h(t,x)$. Let the stochastic processes $(X_\cdot^{t,x})_{x\in\mathbb{R}^d}$ be coupled by sharing the same Brownian motion in their defining SDEs (A.2). The mapping $x \mapsto X_s^{t,x}$ is continuous almost surely as its matrix-valued partial derivative $\partial X_\cdot^{t,x}/\partial x$ solves the SDE

$$\mathrm{d}\frac{\partial X_s^{t,x}}{\partial x} = \nabla\bigl(2b_\infty(X_s^{t,x}) - b_{t-s}(X_s^{t,x})\bigr)\frac{\partial X_s^{t,x}}{\partial x}\,\mathrm{d}s$$

whose wellposedness is guaranteed by the bound

$$|\nabla^2(2b_\infty - b_{t-s})(x)| \leqslant 3 \sup_{m\in\mathcal{P}_2(\mathbb{R}^d)}\sup_{x\in\mathbb{R}}|\nabla^2 D_m F(m,x)| \leqslant C.$$

The norm of $\frac{\partial X_s^{t,x}}{\partial x}$ satisfies

$$\forall s \in [0,t-t_0],\ \forall x \in \mathbb{R}^d, \qquad \left|\frac{\partial X_s^{t,x}}{\partial x}\right| \leqslant C_T \quad \text{a.s.}$$

by Grönwall's lemma. Therefore we have

$$\mathbb{E}\Bigl[\exp\Bigl(C_T^{-1}\sup_{x:|x-x_0|\leqslant 1}|M_{t-t_0}^{t,x}|^2\Bigr)\Bigr] \leqslant C_T\exp(C_T|x_0|^2)$$

for all $x_0 \in \mathbb{R}^d$. We obtain $h(t,x) \to h(t,x_0)$ when $x \to x_0$ by applying the dominated convergence theorem to the Feynman–Kac formula.

We sketch the part for verifying the conditions on derivatives. Differentiate the evolution equation (1.50). We obtain for $k = 1$, 2,

$$\begin{aligned} \partial_t\nabla^k h = \Delta\nabla^k h + (2b_\infty - b_t)\cdot\nabla\nabla^k h + \sum_{i=2}^k \binom{k}{i}\nabla^i(2b_\infty - b_t)\cdot\nabla\nabla^{k-i}h \\ + \sum_{i=1}^k\binom{k}{i}\nabla^i g(t,x)\nabla^{k-i}h + \bigl(\nabla(2b_\infty - b_t)\cdot\nabla\nabla^{k-1}h + g(t,x)\nabla^k h\bigr). \end{aligned}$$

We then write the Feynman–Kac formula for $\nabla^k h$, $k = 1$, 2. The first two terms on the right hand side of the equation corresponds to the same stochastic process, to which the Gaussian moment bound applies. The third and fourth term are lower-order derivatives, continuous in space and have bound $|\nabla^{k-i}h(t,x)| \leqslant \exp\bigl(C_T(1+|x|)\bigr)$ by the induction hypothesis. The last term corresponds to the exponential in the Feynman–Kac formula, whose growth in x remains linear. So we can argue as before to derive $|\nabla^k h(t,x)| \leqslant \exp\bigl(C_T(1+|x|)\bigr)$ for all $(t,x) \in [t_0,T] \times \mathbb{R}^d$. The continuity of $x \mapsto \nabla^k h(t,x)$ for $k=1,2$ follows analogously. Since $x \mapsto h(t,x)$ are twice-differentiable the generalized derivative $\partial_t h$ exists by the evolution equation (1.50). Finally all the constants in the bounds depend only additionally on T, so $(h_t)_{t\in[t_0,T]} \subset \mathcal{A}_+$ uniformly. □

A.2 Proof of modified Bochner's theorem

Proof of Theorem 1.16. We prove the theorem by showing (i) ⇒ (ii) ⇒ (iii) ⇒ (i).

(i) ⇒ (ii). Suppose (i) holds, i.e., $m \mapsto F_{\mathrm{Int}}(m)$ is convex. Let μ be a compactly supported signed measure with $\int \mathrm{d}\mu = 0$. Then it admits decomposition into positive and negative parts: $\mu = \mu_+ - \mu_-$. We define the probability measure

$$m \coloneqq \frac{|\mu|}{\bigl\||\mu|\bigr\|_{\mathrm{TV}}} = \frac{\mu_+ + \mu_-}{\|\mu_+\|_{\mathrm{TV}} + \|\mu_-\|_{\mathrm{TV}}}.$$

Then, for all $t < (\|\mu_+\|_{\mathrm{TV}} + \|\mu_-\|_{\mathrm{TV}})^{-1} \eqqcolon t_0$, we have $m_t \coloneqq m + t\mu \in \mathcal{P}(\mathbb{R}^d)$. Thus, the mapping

$$t \mapsto F_{\mathrm{Int}}(m_t) = F_{\mathrm{Int}}(m) + t\iint V(x-y)m(\mathrm{d}x)\mu(\mathrm{d}y) + \frac{t^2}{2}\iint V(x-y)\mu(\mathrm{d}x)\mu(\mathrm{d}y)$$

is convex on the interval $(-t_0,t_0)$, and therefore, $\iint V(x-y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) \geqslant 0$, which proves (ii).

(ii) ⇒ (iii). Suppose (ii) holds. For non-zero $s \in \mathbb{R}^d$, we define the bounded and continuous function $W_s(t) \coloneqq 2V(t) - V(t+s) - V(t-s)$. Then, for every $\boldsymbol{\xi} \in \mathbb{R}^N$ and every $x^1,\dots,x^N \in \mathbb{R}^d$, we have

$$\begin{aligned}
&\sum_{i,j=1}^N \xi^i\xi^j W_s(x^i-x^j) \\
&\quad= \sum_{i,j=1}^N \xi^i\xi^j V(x^i-x^j) + \sum_{i,j=1}^N \xi^i\xi^j V\bigl((x^i+s-(x^j+s)\bigr) \\
&\qquad\quad - \sum_{i,j=1}^N \xi^i\xi^j V\bigl((x^i+s)-x^j\bigr) - \sum_{i,j=1}^N \xi^i\xi^j V\bigl(x^i-(x^j+s)\bigr) \\
&\quad= \iint V(x-y)\hat\mu(\mathrm{d}x)\hat\mu(\mathrm{d}y) \geqslant 0, \qquad \text{for } \hat\mu = \sum_{i=1}^N \xi^i\delta_{x^i} - \sum_{i=1}^N \xi^i\delta_{x^i+s}
\end{aligned}$$

as the measure $\hat\mu$ has zero net mass. Thus, W_s is a function of positive type, and according to the classical Bochner's theorem [191, Theorem IX.9], its Fourier transform $\hat W_s$ is a positive and finite measure on $\mathbb{R}^d$. On the other hand, denoting by $\hat V$, $\hat W_s$ the Fourier transforms of V, W_s respectively, we have

$$\hat W_s(k) = 2\bigl(1-\cos(k\cdot s)\bigr)\hat V(k)$$

in the sense of tempered distributions. For every $k \neq 0$, we can find a non-zero $s \in \mathbb{R}^d$ such that the mapping $k' \mapsto 1 - \cos(k' \cdot s)$ is lower bounded away from 0 in a neighborhood of k. Thus, in this neighborhood, we have

$$\hat{V}(k') = \frac{\hat{W}_s(k')}{2\big(1 - \cos(k' \cdot s)\big)}.$$

Therefore, the distribution $\hat{V}$ restricted on $\mathbb{R}^d \setminus \{0\}$ is a positive and locally finite measure, which we denote by λ. The difference $\hat{V} - \lambda$, being a Schwartz distribution, is supported on the singleton $\{0\}$, and by the structure theorem (see e.g., [207, Théorème XXXV] and [114, Theorem 2.3.4]), admits decomposition

$$\hat{V} - \lambda = \sum_{|n|=0}^{m} (-1)^{|n|} c_n D^n \delta_0,$$

n being multi-indices, for some $m \in \mathbb{N}$ and $c_n \in \mathbb{C}$. Denote the heat kernel by

$$\rho^\varepsilon(x) = (2\pi\varepsilon)^{-d/2} \exp(-|x|^2/2\varepsilon)$$

and its Fourier transform reads $\hat{\rho}^\varepsilon(k) = (2\pi)^{-d/2} \exp(-2\pi^2 \varepsilon |k|^2)$. Define $V^\varepsilon = V \star \rho^\varepsilon$. We then have

$$\begin{aligned} V^\varepsilon(0) = \langle \rho^\varepsilon, V \rangle = \langle \hat{\rho}^\varepsilon, \hat{V} \rangle &= \Big\langle \hat{\rho}^\varepsilon, \lambda + \sum_{|n|=0}^{m} (-1)^{|n|} c_n D^n \delta_0 \Big\rangle \\ &= \int_{\mathbb{R}^d \setminus \{0\}} \hat{\rho}^\varepsilon \, \mathrm{d}\lambda + \frac{c_0}{(2\pi)^{d/2}} + \sum_{|n|=1}^{m} c_n \nabla^n \hat{\rho}^\varepsilon(0), \end{aligned}$$

where $\langle \hat{\rho}^\varepsilon, \hat{V} \rangle$ is well defined, since $\hat{\rho}^\varepsilon \in \mathcal{S}$ and $\hat{V} \in \mathcal{S}'$. Thanks to the fact that

$$\int_{\mathbb{R}^d \setminus \{0\}} \hat{\rho}^\varepsilon d\lambda \nearrow \lambda(\mathbb{R}^d \setminus \{0\}), \qquad V^\varepsilon(0) \to V(0), \qquad \nabla^n \hat{\rho}^\varepsilon(0) \to 0$$

when $\varepsilon \searrow 0$, for n such that $|n| \geqslant 1$, we can take the limit and obtain that the mass $\lambda(\mathbb{R}^d \setminus \{0\})$ is finite and $c_0 \in \mathbb{R}$. Then the original potential V reads

$$V(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d \setminus \{0\}} e^{ik \cdot x} \lambda(\mathrm{d}k) + \frac{c_0}{(2\pi)^{d/2}} + P(x),$$

where P is an m-th-order polynomial with $P(0) = 0$. The boundedness of V implies that P must be identically zero, which concludes.

(iii) $\Rightarrow$ *(i)*. Suppose (iii) holds. Let μ be an arbitrary signed measure with $\int d\mu = 0$. Then its Fourier transform $\hat{\mu}$ is even, real-valued, belongs to the class C_0 and satisfies $\hat{\mu}(0) = 0$. Thus, we have

$$\begin{aligned} \iint V(x-y) \mu(\mathrm{d}x) \mu(\mathrm{d}y) = \langle V \star \mu, \mu \rangle &= (2\pi)^{d/2} \langle \hat{V} \hat{\mu}, \hat{\mu} \rangle \\ &= (2\pi)^{d/2} \int_{\mathbb{R}^d \setminus \{0\}} \big(\hat{\mu}(k)\big)^2 \hat{V}(\mathrm{d}k) \geqslant 0, \end{aligned}$$

which proves (ii). Finally, from the computation in the first paragraph, we see that (i) is a consequence of (ii). ∎

Appendix B

Appendices to Chapter 2

B.1 Lower-semicontinuities

Lemma B.1. *The entropy* $H : \mathcal{P}_2(\mathbb{R}^{2d}) \to (-\infty, +\infty]$ *and the Fisher information* $I : \mathcal{P}_2(\mathbb{R}^{2d}) \to (-\infty, +\infty]$ *are lower-semicontinuous with respect to the weak topology of* $\mathcal{P}_2$*. Consequently, under the assumption* (2.2)*, if* $(m_n)_{n\in\mathbb{N}}$ *is a sequence converging to* m_* *in* $\mathcal{P}_2(\mathbb{R}^{2d})$*, then*

$$\liminf_{n\to+\infty} H(m_n|\hat{m}_n) \geqslant H(m_*|\hat{m}_*) \quad \text{and} \quad \liminf_{n\to+\infty} I(m_n|\hat{m}_n) \geqslant I(m_*|\hat{m}_*).$$

Proof. The lower semicontinuity of $m \mapsto H(m)$ is classical. We show the lower semicontinuity of the Fisher information. Let $(m_n)_{n\in\mathbb{N}}$ be a sequence converging to m_* in $\mathcal{P}_2(\mathbb{R}^d)$. Without loss of generality, we suppose $I(m_n) \leqslant M^2$ for every $n \in \mathbb{N}$. This implies in particular $\|\nabla m_n\|_{L^1} \leqslant M$ by Cauchy–Schwarz. For every function φ belonging to $C_c^\infty(\mathbb{R}^d)$, we have

$$|\langle \nabla\varphi, m_* \rangle| = \lim_{n\to+\infty} |\langle \nabla\varphi, m_n \rangle| \leqslant M\|\varphi\|_\infty.$$

Hence $\|\nabla m_*\|_{\mathrm{TV}} \leqslant M$ as well. Moreover, for every $f \in C_c(\mathbb{R}^d)$ and every $\varepsilon > 0$, we can find $\varphi \in C_c^\infty(\mathbb{R}^d)$ such that $\|f - \varphi\|_\infty < \frac{\varepsilon}{4M}$ and $n \in \mathbb{N}$ such that $|\langle \nabla\varphi, m_n - m_* \rangle| < \frac{\varepsilon}{2}$. Then,

$$\begin{aligned} |\langle f, \nabla(m_n - m_*) \rangle| &\leqslant |\langle f - \varphi, \nabla m_n \rangle| + |\langle f - \varphi, \nabla m_* \rangle| + |\langle \nabla\varphi, m_n - m_* \rangle| \\ &< 2 \cdot \frac{\varepsilon}{4M} \cdot M + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Equivalently, the sequence of $\mathbb{R}^d$-valued Radon measures $(\nabla m_n)_{n\in\mathbb{N}}$ converges to ∇m *locally* weakly. We then apply [3, Theorem 2.34] to obtain

$$\liminf_{n\to+\infty} I(m_n) \geqslant I(m_*).$$

Finally, the lower semicontinuity of $m \mapsto H(m|\hat{m})$ (resp. $m \mapsto I(m|\hat{m})$) follows from the lower semicontinuity of $m \mapsto H(m)$ (resp. $m \mapsto I(m)$) and the locally uniform quadratic growth of $x \mapsto \frac{\delta F}{\delta m}(m, x)$ (resp. the locally uniform linear growth of $x \mapsto D_m F(m, x)$). ∎

B.2 Convergence of non-linear functional of empirical measures

Let $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a (non-linear) mean field functional and m be a probability measure with finite second moment. We suppose the first and second-order functional derivatives $\frac{\delta\phi}{\delta m}, \frac{\delta^2\phi}{\delta m^2}$ exist and that (ϕ, m) satisfies

$$\forall m' \in \mathcal{P}_2(\mathbb{R}^d),\ \forall x \in \mathbb{R}^d, \quad \big|D_m\phi(m', x)\big| \leqslant M_1, \tag{B.1}$$

$$\forall m' \in \mathcal{P}_2(\mathbb{R}^d), \quad \left|\iint \left[\frac{\delta^2\phi}{\delta m^2}(m', x, x) - \frac{\delta^2\phi}{\delta m^2}(m', x, y)\right] m(dx)m(dy)\right| \leqslant M_2 \tag{B.2}$$

for some constants M_1 and M_2.

Remark B.2. The condition (B.2) is a modified version of the condition [218, (*p*-LFD)]. Our version has the advantage of being *intrinsic*: the left hand side of (B.2) stays invariant under the change $\frac{\delta^2\phi}{\delta m^2}(m, x, y) \to \frac{\delta^2\phi}{\delta m^2}(m, x, y) + \frac{\delta\phi_1}{\delta m}(m, y) + \phi_2(m)$ for regular enough ϕ_1 and ϕ_2.

Lemma B.3. *If the mean field functional ϕ and the measure m satisfy* (B.1) *and* (B.2)*, then for N i.i.d. random variables $\xi_1, \ldots, \xi_N \sim \mu$, we have*

$$\mathbb{E}\big[|\phi(\mu_{\boldsymbol{\xi}}) - \phi(m)|^2\big] \leqslant \frac{M_1^2 \operatorname{Var} m}{N} + \frac{M_2^2}{4N^2}. \tag{B.3}$$

Proof. We have the decomposition

$$\mathbb{E}\big[|\phi(\mu_{\boldsymbol{\xi}}) - \phi(m)|^2\big] = \operatorname{Var}\phi(\mu_{\boldsymbol{\xi}}) + \big(\mathbb{E}\big[\phi(\mu_{\boldsymbol{\xi}})\big] - \phi(m)\big)^2.$$

Thanks to (B.1), the mapping $\xi^i \mapsto \phi(\mu_{\boldsymbol{\xi}})$ is $\frac{M_1}{N}$-Lipschitz continuous, so by the Efron–Stein inequality we have

$$\operatorname{Var}\phi(\mu_{\boldsymbol{\xi}}) \leqslant \frac{M_1^2}{N}\operatorname{Var} m.$$

For the second term we apply the argument of [218, Theorem 4.2.9 (i)] and obtain

$$\big|\mathbb{E}\big[\phi(\mu_{\boldsymbol{\xi}})\big] - \phi(m)\big| \leqslant \frac{M_2}{2N}. \qquad \square$$

B.3 Validity of Girsanov transforms

We prove a lemma similar to [117, Lemma A.1] which allows us to justify Girsanov transforms.

Lemma B.4. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ be a filtered probability space. If β, γ, X, $Y : \Omega \times [0, T] \to \mathbb{R}^d$ are $\mathcal{F}_t$-adapted continuous stochastic processes satisfying*

$$|\beta_t| + |\gamma_t| \leqslant C(1 + |X_t| + |V_t|)$$

almost surely for some constant C, and if the tuple (X, V, β) solves

$$dX_t = V_t dt,$$
$$dV_t = \beta_t dt + \sqrt{2}dW_t$$

for an $\mathcal{F}_t$*-adapted Brownian* W_t *with* $\mathbb{E}\big[|X_0|^2 + |V_0|^2\big] < +\infty$*, then the exponential local martingale*

$$R_\cdot := \exp\bigg(\int_0^\cdot \gamma_s \cdot dW_s - \frac{1}{2}\int_0^\cdot |\gamma_s|^2 ds\bigg)$$

is uniformly integrable.

Proof. It suffices to verify $\mathbb{E}[R_T] = 1$. Put $M_\cdot = R_\cdot(1 + |X_\cdot|^2 + |V_\cdot|^2)$. By Itō's formula, the local semimartingale satisfies

$$dM_t = 2R_t\big(X_t \cdot Y_t + (\beta_t + \sqrt{2}\gamma_t)\cdot Y_t + 1\big)dt + R_t\big((1+|X_t|^2+|Y_t|^2)\gamma_t + 2\sqrt{2}Y_t\big)\cdot dW_t.$$

Using the uniform linear growth condition of β, γ, we can find a constant C such that $t \mapsto e^{-Ct}M_t$ is a local supermartingale. But $e^{-Ct}M_t \geqslant 0$. So by Fatou's lemma $t \mapsto e^{-Ct}M_t$ is really a supermartingale and this yields $\mathbb{E}[M_t] \leqslant e^{Ct}\,\mathbb{E}[M_0]$. The Itō's formula for $R_\cdot$ writes

$$dR_t = R_t\gamma_t \cdot dW_t.$$

So for $\varepsilon > 0$ the bounded supermartingale $\frac{R_\cdot}{1+\varepsilon R_\cdot}$ satisfies

$$d\frac{R_t}{1+\varepsilon R_t} = -\frac{\varepsilon R_t^2\gamma_t^2}{(1+\varepsilon R_t)^3}dt + \frac{R_t}{(1+\varepsilon R_t)^2}\gamma_t \cdot dW_t.$$

Taking expectations on both sides, we obtain

$$\mathbb{E}\bigg[\frac{R_T}{1+\varepsilon R_T}\bigg] = \frac{1}{1+\varepsilon} - \mathbb{E}\bigg[\int_0^T \frac{\varepsilon R_t^2\gamma_t^2}{(1+\varepsilon R_t)^3}dt\bigg].$$

Using the bound $\frac{\varepsilon R_t^2\gamma_t^2}{(1+\varepsilon R_t)^3} \leqslant R_t\gamma_t^2 \leqslant CM_t$, we take the limit $\varepsilon \to 0$ by the dominated convergence theorem and obtain $\mathbb{E}[R_T] = 1$. $\square$

Appendix C

Appendices to Chapter 4

C.1 Well-posedness of singular dynamics

The mean field well-posedness proof will mainly be based on the estimates on the convolution with the kernel K in Proposition 4.15 and the following elementary result.

Proposition C.1 (Growth and stability estimates)**.** *Let $T > 0$ and $\beta : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ be a vector field that is the sum of a Lipschitz and a bounded part, that is, $\beta = \beta_{\mathrm{Lip}} + \beta_{\mathrm{b}}$ with $\nabla\beta_{\mathrm{Lip}}$, $\beta_{\mathrm{b}} \in L^\infty$. Suppose its divergence is lower bounded: $(\nabla \cdot \beta)_- \in L^\infty$. Let $m : [0,T] \to \mathcal{P}(\mathbb{R}^d)$ be a probability solution to the parabolic equation*

$$\partial_t m_t = \Delta m_t - \nabla \cdot (\beta_t m_t) \,.$$

Then, for all $p \in [2,\infty]$, we have

$$\|m_t\|_{L^p} \leqslant C_p\big(\|m_0\|_{L^p} + 1\big)$$

for some C_p depending only on p, d and $\|(\nabla \cdot \beta)_-\|_{L^\infty}$ (notably independent of t and T).

Moreover, let β' be another vector field satisfying the same conditions as β, and let m' be a probability solution to the equation corresponding to β'. Then, for all $p \in \{1\} \cup [2,\infty)$, we have

$$\|m_t - m'_t\|_{L^p} \leqslant e^{C'_p t}\|m_0 - m'_0\|_{L^p} + C'_p\Big((e^{C'_p t} - 1)\mathbb{1}_{p\geqslant 2} + \sqrt{t}\mathbb{1}_{p=1}\Big) \sup_{v\in[0,t]} \|\beta_v - \beta'_v\|_{L^\infty}$$

for some C'_p depending only on p, d, $\|(\nabla \cdot \beta)_-\|_{L^\infty}$, $\|(\nabla \cdot \beta')_-\|_{L^\infty}$, $\|m_0\|_{L^p}$ and $\|m'_0\|_{L^p}$.

Proof. First, consider the SDE

$$\mathrm{d}X_t = \beta_t(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \,.$$

Since its drift is the sum of a bounded and a Lipschitz part, we have the existence of the strong solution and we find that if $\mathrm{Law}(X_0) = m_0$, then we have the correspondence $\mathrm{Law}(X_t) = m_t$, by the uniqueness of the PDE. Moreover, it is known (see e.g. [45]) that if we take a mollified sequence approaching towards β, the SDE solution will also tend to the original one, i.e. X, and we have the continuous dependency on the initial value as well. So without loss of generality, we can suppose that $\nabla\beta \in \mathcal{C}_{\mathrm{b}}^{\infty}$ and m_0 belongs to the Schwartz class. By a Feynman–Kac argument similar to that of Proposition C.2, we know that m_t belongs also to the Schwartz class. Thus, in the following we perform only formal calculations.

Step 1: Growth estimates. Let $p \geqslant 2$. The L^p norm of m_t satisfies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} m_t^p &= p\int_{\mathbb{R}^d} m_t^{p-1}\partial_t m_t \\ &= p\int_{\mathbb{R}^d} m_t^{p-1}\Big(\Delta m_t - \nabla\cdot(\beta_t m_t)\Big) \\ &= \int_{\mathbb{R}^d}\Big(-p(p-1)m_t^{p-2}|\nabla m_t|^2 - (p-1)(\nabla\cdot\beta_t)m_t\Big) \\ &\leqslant -p(p-1)\int_{\mathbb{R}^d} m_t^{p-2}|\nabla m_t|^2 + (p-1)\|(\nabla\cdot\beta)_-\|_{L^\infty}\int_{\mathbb{R}^d} m_t^p \\ &\leqslant (p-1)\|(\nabla\cdot\beta)_-\|_{L^\infty}\int_{\mathbb{R}^d} m_t^p\,,\end{aligned}$$

where here and in the following C_p denotes a constant having the same dependencies as in the statement, and may change from line to line. We would also denote by C a constant that does not depend on p, but having the same other dependencies. By Grönwall's lemma, we get

$$\|m_t\|_{L^p} \leqslant \exp\biggl(\frac{p-1}{p}(\nabla\cdot\beta)_-\|_{L^\infty}t\biggr)\|m_0\|_{L^p}\,, \tag{C.1}$$

and taking $p \to \infty$, we get

$$\|m_t\|_{L^\infty} \leqslant \exp\bigl(\|(\nabla\cdot\beta)_-\|_{L^\infty}t\bigr)\|m_0\|_{L^\infty}\,. \tag{C.2}$$

Now we show that the two estimates above can be improved into time-uniform ones. To this end, define the operator $\mathcal{L}_t = \Delta + \beta_t\cdot\nabla$ and its dual $\mathcal{L}_t^* = \Delta - \nabla\cdot(\beta_t\cdot)$. Denote by $(P_{u,t})_{0\leqslant u\leqslant t\leqslant T}$ the time-dependent semi-group generated. Specializing to $p = 2$ in the L^p computations above, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} m_t^2 \leqslant -2\int_{\mathbb{R}^d}|\nabla m_t|^2 + \|(\nabla\cdot\beta)_-\|_{L^\infty}\int_{\mathbb{R}^d} m_t^2\,.$$

The Nash inequality indicates

$$\|m_t\|_{L^2}^{1+2/d} \leqslant C_d\|m_t\|_{L^1}^{2/d}\|\nabla m_t\|_{L^2}\,,$$

where C_d depends only on d. So by Grönwall's lemma, we get the uniform-in-time bound over $\|m_t\|_{L^2}$:

$$\|m_t\|_{L^2}^2 \leqslant \biggl(\frac{C_d^2\|(\nabla\cdot\beta)_-\|_{L^\infty}}{2}\biggr)^{d/2}(1-e^{-\kappa t})^{-d/2}\|m_t\|_{L^1}^2\,,$$

for $\kappa = 2\|(\nabla\cdot\beta)_-\|_{L^\infty}/d$. Note that this bound is independent of $\|m_0\|_{L^\infty}$. Now we take an arbitrary $h_0 : \mathbb{R}^d \to [0,\infty)$ of the Schwartz class and consider the dual evolution $\partial_u h_u = \mathcal{L}_{t-u}h_u$, that is,

$$\partial_u h_u = \Delta h_u + \beta_{t-u}\cdot\nabla h_u\,,$$

for $u \in [0,t]$, where $t \in [0,T]$. Deriving the L^1 norm of h_s and integrating by parts, we get

$$\|h_u\|_{L^1} \leqslant \exp\big(\|(\nabla\cdot\beta)_-\|_{L^\infty} u\big)\|h_0\|_{L^1}\,.$$

Doing the same for the L^2 norm, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\int_{\mathbb{R}^d} h_u^2 &= 2\int_{\mathbb{R}^d} h_u \mathcal{L}_{T-u} h_u \\ &= 2\int_{\mathbb{R}^d} h_u\big(\Delta h_u + \beta_{t-u}\cdot\nabla h_u\big) \\ &= -2\int_{\mathbb{R}^d}|\nabla h_u|^2 - \int_{\mathbb{R}^d} h_u^2 \nabla\cdot\beta_{t-u} \\ &\leqslant -2\int_{\mathbb{R}^d}|\nabla h_u|^2 + \|(\nabla\cdot\beta)_-\|_{L^\infty}\int_{\mathbb{R}^d} h_u^2\,.\end{aligned}$$

Again, using the Nash inequality:

$$\|h_u\|_{L^2}^{1+2/d} \leqslant C_d\|h_u\|_{L^1}^{2/d}\|\nabla h_u\|_{L^2} \leqslant C_d \exp\big(2\|(\nabla\cdot\beta)_-\|_{L^\infty}u/d\big)\|h_0\|_{L^1}^{2/d}\|\nabla h_u\|_{L^2}\,,$$

we derive the bound over $\|h_t\|_{L^2}$:

$$\|P_{t-u,t}h_0\|_{L^2}^2 = \|h_u\|_{L^2}^2 \leqslant \left(\frac{C_d^2\|(\nabla\cdot\beta)_-\|_{L^\infty}}{2}\right)^{d/2}(e^{-\kappa u} - e^{-2\kappa u})^{-d/2}\|h_0\|_{L^1}^2\,,$$

from which follows the bound on $\|P_{t-u,t}\|_{L^1\to L^2}$. So, taking $u = \max(t/2, t-\kappa^{-1})$, we get

$$\begin{aligned}\|m_t\|_{L^\infty} = \|P_{u,t}^* m_u\|_{L^\infty} &\leqslant \|m_u\|_{L^2}\|P_{u,t}^*\|_{L^2\to L^\infty} \\ &= \|m_u\|_{L^2}\|P_{u,t}\|_{L^1\to L^2} \leqslant C(t\wedge 1)^{-d/2}\|m_0\|_{L^1}\,. \end{aligned} \tag{C.3}$$

So, combining (C.2) and (C.3), we get a uniform-in-time bound over $\|m_t\|_{L^\infty}$:

$$\sup_{t\in[0,T]}\|m_t\|_{L^\infty} \leqslant C\big(\|m_0\|_{L^1} + \|m_0\|_{L^\infty}\big)\,. \tag{C.4}$$

Finally, by differentiating $\int m_t$ and integrating by parts, we get

$$\|m_t\|_{L^1} = \|m_0\|_{L^1}\,. \tag{C.5}$$

Similarly, interpolating between (C.3) and (C.5), we get

$$\|m_t\|_{L^p} \leqslant C^{(p-1)/p}(t\wedge 1)^{-(p-1)d/2p}\|m_0\|_{L^1}\,,$$

and combing with (C.1), we get

$$\sup_{t\in[0,T]}\|m_t\|_{L^p} \leqslant C_p\big(\|m_0\|_{L^1} + \|m_0\|_{L^p}\big)\,. \tag{C.6}$$

Step 2: Stability estimates. Now let β', m' be the other vector field and the probability solution. Recall that m, m' correspond respectively to the SDE

$$\begin{aligned}\mathrm{d}X_t &= \beta_t(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t\,, &\quad \mathrm{Law}(X_0) = m_0\,,\\ \mathrm{d}X_t' &= \beta_t'(X_t')\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t\,, &\quad \mathrm{Law}(X_0') = m_0'\,.\end{aligned}$$

Now introduce the third SDE, whose drift term is identical to the first, but initial condition identical to the second:

$$\mathrm{d}X''_t = \beta_t(X''_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t\,, \qquad \mathrm{Law}(X''_0) = m'_0\,.$$

such that $\mathbb{P}[X_0 \neq X''_0] = \frac{1}{2}\|m_0 - m'_0\|_{L^1}$. Denote $m''_t = \mathrm{Law}(X''_t)$. Thus, conditioning on the initial condition, we get

$$\|m_t - m''_t\|_{L^1} \leqslant 2\,\mathbb{P}[X_t \neq X''_t] \leqslant 2\,\mathbb{P}[X_0 \neq X''_0] = \|m_0 - m'_0\|_{L^1}\,.$$

On the other hand, by Pinsker's inequality and Girsanov's theorem, we have

$$\|m'_t - m''_t\|^2_{L^1} \leqslant 2H(m'_t|m''_t) \leqslant \frac{1}{2}\int_0^t \|\beta_v - \beta'_v\|^2_{L^\infty}\,\mathrm{d}v\,.$$

Combining the two inequalities above yields the L^1-stability estimate. Now, let $p \geqslant 2$ and let us calculate:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}|m_t - m'_t|^p \\
&\quad= p\int_{\mathbb{R}^d}|m_t - m'_t|^{p-2}(m_t - m'_t)\Big(\Delta(m_t - m'_t) - \nabla\cdot(m_t\beta_t) + \nabla\cdot(m'_t\beta'_t)\Big) \\
&\quad= -p(p-1)\int_{\mathbb{R}^d}|m_t - m'_t|^{p-2}|\nabla(m_t - m'_t)|^2 \\
&\qquad\quad + p\int_{\mathbb{R}^d}|m_t - m'_t|^{p-2}(m_t - m'_t)\Big(-\nabla\cdot(m_t\beta_t) + \nabla\cdot(m'_t\beta'_t)\Big) \\
&\quad= -p(p-1)\int_{\mathbb{R}^d}|m_t - m'_t|^{p-2}|\nabla(m_t - m'_t)|^2 \\
&\qquad\quad + p\int_{\mathbb{R}^d}|m_t - m'_t|^{p-2}(m_t - m'_t)\Big(-\nabla\cdot\big((m_t - m'_t)\beta'_t\big) + \nabla\cdot\big(m_t(\beta'_t - \beta_t)\big)\Big) \\
&\quad\leqslant (p-1)\int_{\mathbb{R}^d}|m_t - m'_t|^p\cdot(-\nabla\cdot\beta'_t) + \frac{p(p-1)}{4}\int_{\mathbb{R}^d}|m_t - m'_t|^{p-2}m_t^2|\beta_t - \beta'_t|^2 \\
&\quad\leqslant \frac{(p-1)}{2}\|(\nabla\cdot\beta')_-\|_{L^\infty}\int_{\mathbb{R}^d}|m_t - m'_t|^p + \frac{(p-1)(p-2)}{4}\int_{\mathbb{R}^d}|m_t - m'_t|^p \\
&\qquad\quad + \frac{p-1}{2}\|m_t\|^p_{L^p}\|\beta_t - \beta'_t\|^p_{L^\infty}\,.
\end{aligned}$$

Then, using the uniform L^p estimate in the first step, applying Grönwall's lemma and taking the p-th root, we get the desired result. □

Now we are ready to prove the well-posedness of the mean field dynamics.

Proof of Proposition 4.16. Take a $p \in (1,\infty)$ such that $p^{-1} < 1 - \frac{s+1}{d}$ and let q be its conjugate: $p^{-1} + q^{-1} = 1$. We also take a $\theta \in (0, d - s - 1)$.

Step 1: Well-posedness. Let $T \in (0,\infty)$. We define the functional space:

$$\mathcal{X} \coloneqq \mathcal{C}([0,T]; L^1 \cap L^p \cap \mathcal{P})\,.$$

The space $\mathcal{X}$ is a complete metric space. Given $m \in X$, we let $\mathcal{T}[m]$ be the uniqueness probability solution to the Cauchy problem

$$\partial_t\mathcal{T}[m]_t = \Delta\mathcal{T}[m]_t - \nabla\cdot\big((K\star m_t - \nabla U)\mathcal{T}[m]_t\big)\,, \quad \mathcal{T}[m]_0 = m_0\,.$$

According to Proposition C.1 we know that $\mathcal{T}[m] \in \mathcal{X}$, where the continuity in $L^1 \cap L^p$ follows from a density argument. Moreover, by the stability estimate in the proposition, for all m, $m' \in X$, we have

$$\|\mathcal{T}[m]_t - \mathcal{T}[m']_t\|_{L^r} \leqslant e^{C'_r t}\|m_0 - m'_0\|_{L^r} + C'_r\Big((e^{C'_r t} - 1)\mathbb{1}_{r=p} + \sqrt{t}\mathbb{1}_{r=1}\Big) \sup_{v\in[0,t]} \|K \star (m_v - m'_v)\|_{L^\infty}$$

for $r = 1$, p. But by Proposition 4.15, we have

$$\|K \star (m_v - m'_v)\|_{L^\infty} \lesssim \|m_v - m'_v\|_{L^1}^{1-q(s+1)/d}\|m_v - m'_v\|_{L^p}^{q(s+1)/d}.$$

Thus, restricting to the subspace of $\mathcal{X}$ of common initial value and letting T be small enough, we get that the mapping $\mathcal{T}$ is a contraction in $\mathcal{X}$. So a time-local solution exists and is unique. Thanks to the uniform growth estimates, this short time interval can be extended infinitely by iteration. So a unique global solution is recovered and it satisfies the uniform L^∞ bound thanks to Proposition C.1. For the continuous dependency on the initial value, we use the stability estimates on a small time interval without restricting the initial values to be the same and iterate infinitely as well.

Step 2: Control of moments. Given the uniform L^∞ bound obtained above, we have, according to Proposition 4.15,

$$\|K \star m_t\|_{L^\infty} \lesssim \|m_t\|_{L^1}^{1-(s+1)/d}\|m_t\|_{L^\infty}^{(s+1)/d}.$$

So the contribution from the interaction kernel is bounded. Then we construct, for $k > 0$, the Lyapunov function

$$V_k(x) = \sqrt{1 + |x|^{2k}}\,,$$

and we can easily verify

$$\big(\Delta - \nabla U \cdot \nabla + (K \star m_t) \cdot \nabla\big)V_k \leqslant -c_k V_k + C_k\,,$$

for some $c_k > 0$, $C_k \geqslant 0$. This implies the uniform bound on the k-th moment.

Step 3: Approximation. Let $(m^\varepsilon_t)_{t\geqslant 0}$ be the flow corresponding to the mollified kernel K^ε and potential U^ε. Applying the stability estimates in Proposition C.1, we get

$$\|m_t - m^\varepsilon_t\|_{L^r} \leqslant e^{C'_r t}\|m_0 - m^\varepsilon_0\|_{L^r} + C'_r\Big((e^{C'_r t} - 1)\mathbb{1}_{r=p} + \sqrt{t}\mathbb{1}_{r=1}\Big) \sup_{v\in[0,t]} \Big(\|K \star m_v - K^\varepsilon \star m^\varepsilon_v\|_{L^\infty} + \|\nabla U - \nabla U^\varepsilon\|_{L^\infty}\Big).$$

Note that the initial L^p error $\|m_t - m^\varepsilon_t\|_{L^p} \to 0$ by interpolation between L^1 and L^∞. For the first term in the supremem, we have

$$\begin{aligned}\|K \star m_v - K^\varepsilon \star m^\varepsilon_v\|_{L^\infty} &\leqslant \|K \star (m_v - m^\varepsilon_v)\|_{L^\infty} + \|K \star m^\varepsilon_v - K \star m^\varepsilon_v \star \eta^\varepsilon\|_{L^\infty} \\ &\leqslant \|K \star (m_v - m^\varepsilon_v)\|_{L^\infty} + \varepsilon^\theta [K \star m^\varepsilon_v]_{\mathcal{C}^\theta}\,.\end{aligned}$$

By the L^∞ and Hölder estimates in Proposition 4.15, we have the following controls:

$$\|K \star (m_v - m_v^\varepsilon)\|_{L^\infty} \lesssim \|m_v - m_v^\varepsilon\|_{L^1}^{1-q(s+1)/d} \|m_v - m_v^\varepsilon\|_{L^p}^{q(s+1)/d},$$
$$[K \star m_v^\varepsilon]_{\mathcal{C}^\theta} \lesssim \|m_v^\varepsilon\|_{L^1}^{1-(s+1+\theta)/d} \|m_v^\varepsilon\|_{L^\infty}^{(s+1+\theta)/d}.$$

For the second term we simply bound $\|\nabla U - \nabla U^\varepsilon\|_{L^\infty} \leqslant \|\nabla^2 U\|_{L^\infty}\varepsilon$. Since m^ε is again uniformly bounded in $L^1 \cap L^\infty$, we get an error bound between m_t and m_t^ε for small t and we iterate infinitely. □

Finally, we prove the well-posedness of the particle system in the non-attractive sub-Coulombic and Coulombic cases.

Proof of Proposition 4.17. Define for $n \in \mathbb{N}$ the sequence of stopping times:

$$\tau_n \coloneqq \inf\bigl\{t \geqslant 0 : \bigl|X_t^i - X_t^j\bigr| \leqslant 1/n \text{ for some } i \neq j\bigr\}.$$

Then the original SDE system (4.39) stopped at τ_n is well defined according to Cauchy–Lipschitz theory. Consider the "energy" functional

$$E(\boldsymbol{x}) = E(x^1, \dots, x^N) = \frac{1}{2} \sum_{\substack{i,j \in [\![1,N]\!] \\ i \neq j}} g_s(x^i - x^j) + \frac{N \mathbb{1}_{s=0}}{2} \sum_{i=1}^N |x^i|^2.$$

The energy functional is always lower bounded, and by Itō calculus, we find that $\mathbb{E}\bigl[E\bigl(\boldsymbol{X}_{t\wedge\tau_n}\bigr)\bigr]$ is upper bounded uniformly in n. Then using the Markov inequality for the energy, we show that $\mathbb{P}[\tau_n \leqslant t] \to 0$ when $n \to \infty$. This implies that $\lim_{n\to\infty} \tau_n = \infty$ almost surely, thus the local well-posedness of the SDE extends to the half line $[0, \infty)$. That is to say the first claim is proved.

Now prove the second claim. For each $n \in \mathbb{N}$, we construct a Lipschitz kernel $\tilde{K}_n : \mathbb{R}^d \to \mathbb{R}$ such that $\tilde{K}_n(x) = K(x)$ for $x \in \mathbb{R}^d$ with $|x| \geqslant 1/n$. Define the convolution $\tilde{K}_n^\varepsilon = \tilde{K}_n \star \eta^\varepsilon$ and consider the SDE system

$$\mathrm{d}\tilde{X}_{n,t}^{\varepsilon,i} = -\nabla U^\varepsilon\bigl(\tilde{X}_{n,t}^{\varepsilon,i}\bigr)\,\mathrm{d}t + \frac{1}{N-1} \sum_{j \in [\![1,N]\!] \setminus \{i\}} \tilde{K}_n^\varepsilon\bigl(\tilde{X}_{n,t}^{\varepsilon,i} - \tilde{X}_{n,t}^{\varepsilon,j}\bigr)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i,$$

for $i \in [\![1, N]\!]$, with initial condition $\tilde{\boldsymbol{X}}_{n,0}^\varepsilon = \boldsymbol{X}_0$. Define the stopping time

$$\tau_n^\varepsilon \coloneqq \inf\bigl\{t \geqslant 0 : \bigl|\tilde{X}_{n,t}^{\varepsilon,i} - \tilde{X}_{n,t}^{\varepsilon,j}\bigr| \leqslant 1/n + \varepsilon \text{ for some } i \neq j\bigr\}.$$

By construction, we know

$$\tilde{\boldsymbol{X}}_{n,t\wedge\tau_n^\varepsilon}^\varepsilon = \boldsymbol{X}_{t\wedge\tau_n^\varepsilon}^\varepsilon \text{ a.s.}$$

On the other hand, by Cauchy–Lipschitz theory, we know

$$\sup_{t\in[0,T]} \bigl|\tilde{\boldsymbol{X}}_{n,t\wedge\tau_n}^\varepsilon - \boldsymbol{X}_{t\wedge\tau_n}\bigr| \leqslant C(n, N, K, U, T)\varepsilon \text{ a.s.}$$

Thus, for each $n \in \mathbb{N}$, there exists $\varepsilon_0(n, N, K, U, T) > 0$ such that for all $\varepsilon \leqslant \varepsilon_0$, we have

$$\sup_{t\in[0,T]} \bigl|\tilde{\boldsymbol{X}}_{n,t\wedge\tau_n}^\varepsilon - \boldsymbol{X}_{t\wedge\tau_n}\bigr| \leqslant \frac{1}{3n} \text{ a.s.}$$

In particular, we get for all $\varepsilon \leqslant \varepsilon_0$, $t \leqslant T \wedge \tau_n$ and $i \neq j$,

$$\left|\tilde{X}^{\varepsilon,i}_{n,t} - \tilde{X}^{\varepsilon,j}_{n,t}\right| \geqslant \frac{1}{3n} \text{ a.s.}$$

Consequently, for $\varepsilon \leqslant \varepsilon_1(n, N, K, U, T) \coloneqq \varepsilon_0 \wedge 1/(13n)$, we have $T \wedge \tau_n \leqslant \tau^{\varepsilon}_{4n}$, and therefore,

$$\begin{aligned}\sup_{t\leqslant T\wedge\tau_n} \left|\boldsymbol{X}^{\varepsilon}_t - \boldsymbol{X}_t\right| = \sup_{t\leqslant T\wedge\tau_n} \left|\boldsymbol{X}^{\varepsilon}_{t\wedge\tau^{\varepsilon}_{4n}} - \boldsymbol{X}_t\right| = \sup_{t\leqslant T\wedge\tau_n} \left|\tilde{\boldsymbol{X}}^{\varepsilon}_{4n,t} - \boldsymbol{X}_t\right| \\ \leqslant C(4n, N, K, U, T)\varepsilon \text{ a.s.}\end{aligned}$$

Thus, taking $\varepsilon \to 0$, we get $\boldsymbol{X}^{\varepsilon}_{t\wedge\tau_n} \to \boldsymbol{X}_{t\wedge\tau_n}$ a.s. for all $t \leqslant T$. We recover the second claim by using the arbitrariness of T and the fact that $\lim_{n\to\infty} \tau_n = \infty$ a.s. □

C.2 Feynman–Kac formula

Proposition C.2. *Let $T > 0$. Suppose $\beta : [0, T]\times\mathbb{R}^d \to \mathbb{R}^d$ and $\varphi : [0, T]\times\mathbb{R}^d \to \mathbb{R}$ are measurable functions and suppose that there exists $C > 0$ such that for all $t \in [0, T]$ and $x \in \mathbb{R}^d$, we have*

$$\begin{aligned} |\beta(t, x)| &\leqslant C(1 + |x|)\,, \\ |\varphi(t, x)| &\leqslant C(1 + |x|)\,, \\ \left|\nabla^k_x \beta(t, x)\right| &\leqslant C\,, \quad \textit{for } k \in [\![1, 3]\!]\,, \\ \left|\nabla^k_x \varphi(t, x)\right| &\leqslant C\,, \quad \textit{for } k \in [\![1, 2]\!]\,. \end{aligned}$$

Suppose in addition that $f_0 : \mathbb{R}^d \to \mathbb{R}$ is measurable and satisfies, for the same constant C, and for all $x \in \mathbb{R}^d$,

$$|\nabla^k f_0(x)| \leqslant C \exp\bigl(C(1 + |x|)\bigr)\,, \quad \textit{for } k \in [\![0, 2]\!]\,.$$

Then, the function $f : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(t, x) = \mathbb{E}\left[\exp\left(\int_0^t \varphi\bigl(t - u, X^{t,x}_u\bigr)\,\mathrm{d}u\right) f_0\bigl(X^{t,x}_t\bigr)\right],$$

where $X^{t,x}_\cdot$ solves

$$\mathrm{d}X^{t,x}_u = \beta\bigl(t - s, X^{t,x}_u\bigr)\,\mathrm{d}u + \sqrt{2}\,\mathrm{d}B_t\,, \quad u \in [0, t]\,, \qquad X^{t,x}_0 = x\,,$$

is a strong solution to the Cauchy problem

$$\partial_t f = \Delta f + \beta \cdot \nabla f + \varphi f\,, \qquad f|_{t=0} = f_0$$

with the following bound: there exists $C' > 0$ such that for all $t \in [0, T]$ and $x \in \mathbb{R}^d$, we have

$$|\nabla^k f_0(t, x)| \leqslant C' \exp\bigl(C'(1 + |x|)\bigr)\,, \quad \textit{for } k \in [\![0, 2]\!]\,.$$

The result can be easily obtained by differentiating the defining SDE of the process $X^{t,x}_\cdot$. We refer readers to e.g. Appendix A.1 for details.

Appendix D

Appendices to Chapter 7

D.1 A lemma on Wasserstein duality

Lemma D.1. *Let $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ be a $\mathcal{C}^2$ function such that the Euclidean operator norm of $\nabla_x \nabla_y f(x, y)$ satisfies*

$$|\nabla_x \nabla_y f(x, y)| \leqslant M$$

for all $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$. Then, for all μ, $\mu' \in \mathcal{P}_2(\mathbb{R}^{d_1})$ and ν, $\nu' \in \mathcal{P}_2(\mathbb{R}^{d_2})$, we have

$$\left| \iint f(x, y)(\mu - \mu')(\mathrm{d}x)(\nu - \nu')(\mathrm{d}y) \right| \leqslant M W_2(\mu, \mu') W_2(\nu, \nu').$$

Proof of Lemma D.1. Let $\pi \in \Pi(\mu, \mu')$ be the W_2-optimal transport plan between μ and μ', and $\pi' \in \Pi(\nu, \nu')$ be that between ν and ν'. Construct the random variable $\big((X, X'), (Y, Y')\big)$ distributed as $\pi \otimes \pi'$. Then, we have

$$\begin{aligned}
&\iint f(x, y)(\mu - \mu')(\mathrm{d}x)(\nu - \nu')(\mathrm{d}y) \\
&\quad = \mathbb{E}[f(X, Y) - f(X', Y) - f(X, Y') + f(X', Y')] \\
&\quad = \mathbb{E}\bigg[\iint_{[0,1]^2} (X - X')^\top \\
&\qquad\qquad \nabla_x \nabla_y f\big((1 - t)X + tX', (1 - s)Y + sY'\big)(Y - Y') \,\mathrm{d}t \,\mathrm{d}s\bigg].
\end{aligned}$$

Therefore, taking absolute values, we get

$$\begin{aligned}
&\left| \iint f(x, y)(\mu - \mu')(\mathrm{d}x)(\nu - \nu')(\mathrm{d}y) \right| \\
&\quad \leqslant M \,\mathbb{E}\big[|X - X'||Y - Y'|\big] \\
&\quad \leqslant M \,\mathbb{E}\big[|X - X'|^2\big]^{1/2} \,\mathbb{E}\big[|Y - Y'|^2\big]^{1/2} \\
&\quad = M W_2(\mu, \mu') W_2(\nu, \nu'),
\end{aligned}$$

which concludes the proof. □

D.2 Algorithm

Algorithm 5: Training two-layer neural network by self-interacting diffusions

Input : Activation func. φ, truncation func. τ, dataset $\{z_k, y_k\}_{k=1}^K$, volatility σ, regularization constant γ, initial distribution m_0, time step δt, time horizon T, non-increasing piecewise constant func. $\lambda : [0, T] \to (0, \infty)$

Output: Y_T

generate $X_0 = (C_0, A_0, B_0) \sim m_0$; $Y_0 = l(X_0)$;

for $t = 0,\ \delta t,\ 2\delta t,\ \dots,\ T - \delta t$ **do**

- generate i.i.d. $\mathcal{N}_t \sim \mathcal{N}(0, 1)$;
- $X_{t+\delta t} \leftarrow X_t - \big(\frac{1}{K}\sum_{k=1}^K Y_t^k \nabla f(X_t, z_k) + \sigma^2\gamma X_t\big)\delta t + \sigma\sqrt{\delta t}\mathcal{N}_t$;
- **if** $\lambda(t) \equiv \lambda_*$ *is constant function* **then**
  - $Y_{t+\delta t} \leftarrow (1 - \lambda_*\delta t)Y_t + \lambda_*\delta t\, \ell(X_{t+\delta t})$;
  - `// Update corresponding to fixed value of` λ
- **else**
  - $Y_{t+\delta t} \leftarrow \big(1 - \lambda(t)\delta t\big)Y_t + \lambda(t)\delta t\, \ell(X_{t+\delta t})$;
  - `// Update corresponding to annealing` $\lambda(t) = \lambda_{\mathrm{a}}(t)$

Appendix E

Appendices to Chapter 8

E.1 Regularity of solution to HJB equation

Throughout this section, we assume that Assumptions 8.5, 8.6 and 8.16 hold and we fix a time horizon $T < +\infty$. Let u be the unique viscosity solution to the HJB equation (8.20). We start by establishing upper and lower bounds on u.

Lemma E.1. *It holds for all $t \in [0, T]$, $x \in \mathbb{R}^d$,*

$$-C_T \leqslant u(t, x) \leqslant C_T(1 + |x|^2).$$

Proof. Under Assumption 8.5 and 8.16, we have $-C_T \leqslant \frac{\delta F}{\delta p}(m_t, x) \leqslant C_T(1 + |x|^2)$. Additionally, under Assumption 8.6, the initial value satisfies $-C \leqslant u_0(x) \leqslant C(1 + |x|^2)$. The desired result follows from the comparison principle. □

To show the existence and uniqueness of the classical solutions to HJB equation (8.20), it is convenient to consider the change of variable $\psi := e^{-u/2}$, which corresponds to the well-known Cole–Hopf transformation.

Lemma E.2. *The function ψ is the unique viscosity solution to*

$$\partial_t \psi_t = \frac{\sigma^2}{2} \Delta \psi_t - \frac{1}{2}\left(\frac{\delta F}{\delta p}(m_t, \cdot) - \gamma u_t\right)\psi_t, \qquad \psi_0(x) = \exp\bigl(-u_0(x)/2\bigr). \tag{E.1}$$

Moreover, it admits the following probabilistic representation

$$\psi(t, x) = \mathbb{E}\left[\exp\left(-\frac{1}{2}\int_0^t \left(\frac{\delta F}{\delta p}(m_{t-s}, x+\sigma W_s) - \gamma u(t-s, x+\sigma W_s)\right) \mathrm{d}s\right)\psi_0(x+\sigma W_t)\right]. \tag{E.2}$$

Proof. First, it follows from the monotonicity of $x \mapsto e^{-x/2}$ that ψ is a viscosity solution to (E.1) if and only if u is a viscosity solution to (8.20). Then, by the bound of u in Lemma E.1, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\gamma}{2}\int_0^t u(t-s, x+\sigma W_s)\,\mathrm{d}s\right)\right] \\
&\quad \leqslant \frac{1}{t}\int_0^t \mathbb{E}\left[\exp\left(\frac{\gamma t}{2} u(t-s, x+\sigma W_s)\right)\right]\mathrm{d}s \\
&\quad \leqslant \frac{1}{t}\int_0^t \mathbb{E}\left[\exp\left(\frac{\gamma t C_T}{2}(1 + |x+\sigma W_s|^2)\right)\right]\mathrm{d}s < \infty,
\end{aligned}$$

for all $t \leqslant \delta$ with δ small enough. Also note that $\big(-\frac{\delta F}{\delta p}(m_t,\cdot)\big)_{t\in[0,T]}$ and ψ_0 are bounded from above. So for $t \leqslant \delta$ we may define

$$\tilde{\psi}(t,x) := \mathbb{E}\bigg[\exp\bigg(-\frac{1}{2}\int_0^t \Big(\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma u(t-s,x+\sigma W_s)\Big)\,\mathrm{d}s\bigg)\psi_0(x+\sigma W_t)\bigg].$$

It is easy to verify that $\tilde{\psi}$ is a viscosity solution to (E.1), so equal to ψ on $[0,\delta]$. Also note that $\psi = \exp(-u/2) \leqslant C_T$ thanks to Lemma E.1. So we may further define for $t \in (\delta, 2\delta]$,

$$\begin{aligned}
&\psi(t,x)\\
&:= \mathbb{E}\bigg[\exp\bigg(-\frac{1}{2}\int_0^{t-\delta} \Big(\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma u(t-s,x+\sigma W_s)\Big)\,\mathrm{d}s\bigg)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\tilde{\psi}(\delta,x+\sigma W_{t-\delta})\bigg]\\
&= \mathbb{E}\bigg[\exp\bigg(-\frac{1}{2}\int_0^{t} \Big(\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma u(t-s,x+\sigma W_s)\Big)\,\mathrm{d}s\bigg)\psi_0(x+\sigma W_t)\bigg].
\end{aligned}$$

Therefore the desired probabilistic representation (E.2) follows from induction. □

Proposition E.3. *The function $\psi = \exp(-u/2) \in \mathcal{C}^3(Q_T)\cap\mathcal{C}(\bar{Q}_T)$ is the unique classical solution to (E.1). Moreover, the gradient $\nabla\psi$ satisfies the growth condition $|\nabla\psi(t,x)| \leqslant C_T(1+|x|^2)$.*

Proof. It follows from Lemma E.2 that

$$s \mapsto \exp\bigg(-\frac{1}{2}\int_0^s \Big(\frac{\delta F}{\delta p}(m_{t-r},x+\sigma W_r)-\gamma u(t-r,x+\sigma W_r)\Big)\,\mathrm{d}r\bigg)\psi(t-s,x+\sigma W_s)$$

is a continuous martingale on $[0,t]$. By Itô's formula, we have for all $0 \leqslant r \leqslant t$ that

$$\begin{aligned}
\psi(t,x) = \mathbb{E}[\psi(r,x+\sigma W_{t-r})]&\\
-\frac{1}{2}\,\mathbb{E}\bigg[\int_0^{t-r} &\Big(\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma u(t-s,x+\sigma W_s)\Big)\\
&\qquad\qquad\psi(t-s,x+\sigma W_s)\,\mathrm{d}s\bigg].
\end{aligned}\tag{E.3}$$

Recall that $\big|\frac{\delta F}{\delta p}(m_t,x)\big| + |u(t,x)| \leqslant C_T(1+|x|^2)$ on $[0,T]\times\mathbb{R}^d$, so for all $t \leqslant T$ we have

$$\int_0^t \mathbb{E}\bigg[\bigg|\Big(\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma u(t-s,x+\sigma W_s)\Big)\psi(t-s,x+\sigma W_s)\frac{W_s}{\sigma s}\bigg|\bigg]\,\mathrm{d}s < \infty.$$

As a result $\nabla\psi$ exists and is equal to

$$\begin{aligned}
\nabla\psi(t,x) = \mathbb{E}[\nabla\psi_0(x+\sigma W_t)]&\\
-\frac{1}{2}\,\mathbb{E}\bigg[\int_0^{t} &\Big(\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma u(t-s,x+\sigma W_s)\Big)\\
&\qquad\qquad\psi(t-s,x+\sigma W_s)\frac{W_s}{\sigma s}\,\mathrm{d}s\bigg].
\end{aligned}$$

Therefore we obtain $|\nabla\psi(t,x)| \leqslant C_T(1+|x|^2)$, and

$$|\nabla u(t,x)| = 2\left|\frac{\nabla\psi(t,x)}{\psi(t,x)}\right| \leqslant C_T(1+|x|^2)\exp\big(C_T(1+|x|^2)\big).$$

In particular we have $\mathbb{E}\big[|\nabla u(t,x+\sigma W_s)|^2\big] < \infty$ for s small enough. So for $r<t$ and r close enough to t we have

$$\begin{aligned}\nabla\psi(t,x) = \mathbb{E}[\nabla\psi(r,x+\sigma W_{t-r})]\\ -\frac{1}{2}\,\mathbb{E}\Big[\int_0^{t-r}\Big(\nabla\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s) - \gamma\nabla u(t-s,x+\sigma W_s)\Big)\\ \psi(t-s,x+\sigma W_s)\,\mathrm{d}s\Big].\end{aligned}$$

Further note that

$$\int_0^{t-r}\mathbb{E}\Big[\Big|\Big(\nabla\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma\nabla u(t-s,x+\sigma W_s)\Big)\psi(t-s,x+\sigma W_s)\frac{W_s}{\sigma s}\Big|\Big]\,\mathrm{d}s < \infty,$$

So $\nabla^2\psi$ exist and is equal to

$$\begin{aligned}\nabla^2\psi(t,x) = \mathbb{E}\Big[\nabla\psi(r,x+\sigma W_{t-r})\frac{W_{t-r}}{\sigma(t-r)}\Big]\\ -\frac{1}{2}\,\mathbb{E}\Big[\int_0^{t-r}\Big(\nabla\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s) - \gamma\nabla u(t-s,x+\sigma W_s)\Big)\\ \psi(t-s,x+\sigma W_s)\frac{W_s}{\sigma s}\,\mathrm{d}s\Big].\end{aligned}$$

Further, in order to compute the time partial derivative, recall (E.3). Since we have already proved that $x\mapsto\psi(t,x)$ belongs to $\mathcal{C}^2$, it follows from Itô's formula that

$$\begin{aligned}&\psi(t,x)-\psi(r,x)\\ &= \frac{\sigma^2}{2}\mathbb{E}\Big[\int_0^{t-r}\Delta\psi(r,x+\sigma W_s)\,\mathrm{d}s\Big]\\ &\quad -\frac{1}{2}\,\mathbb{E}\Big[\int_0^{t-r}\Big(\frac{\delta F}{\delta p}(m_{t-s},x+\sigma W_s)-\gamma u(t-s,x+\sigma W_s)\Big)\psi(t-s,x+\sigma W_s)\,\mathrm{d}s\Big].\end{aligned}$$

Then clearly $\partial_t\psi$ exists and ψ satisfies (E.1). in the classical sense. Moreover, using the same argument, we can easily show that $\nabla^3\psi$ and $\partial_t\nabla\psi$ exist and are continuous on Q_T. □

E.2 Gaussian bounds

The aim of this section is to establish a technical result which ensures that if a family of probability distributions writes as the exponential of a sum of a Lipschitz and a convex function then it admits uniform Gaussian bounds.

Lemma E.4. *Let $p = \exp(-v-w)$ be a probability measure on $\mathbb{R}^d$ that satisfies the following conditions:*

(i) *For some* $\bar{\eta} > \underline{\eta} > 0$*, it holds* $\underline{\eta} I_d \leqslant \nabla^2 v \leqslant \bar{\eta} I_d$*.*

(ii) *The vector* $\nabla v(0)$ *is bounded by* C_1*, i.e.,* $|\nabla v_t(0)| \leqslant C_1$*.*

(iii) *The gradient* ∇w *is bounded by* C_2*, i.e.,* $\|\nabla w_t\|_\infty \leqslant C_2$*.*

Then there exist $\underline{c}$*,* $\bar{c}$*,* $\underline{C}$*,* $\bar{C} > 0$*, depending only on the constants in the conditions and the dimension* d*, such that* $x \in \mathbb{R}^d$*,*

$$\underline{C} \exp(-\underline{c}|x|^2) \leqslant p_t(x) \leqslant \bar{C} \exp(-\bar{c}|x|^2).$$

Proof. We decompose the probability measure $p = qr$ with $q = \exp(-v)/\int \exp(-v)$ and $r = \exp(-w)/\int \exp(-w)q$.

Step 1. We first derive some estimates on v and the corresponding measure q. From Assumption (i), the following inequalities holds

$$|\nabla v(x) - \nabla v(0)||x| \geqslant \big(\nabla v(x) - \nabla v(0)\big) \cdot x \geqslant \underline{\eta}|x|^2$$

Let x_* be the unique solution to $\nabla v(x) = 0$, i.e., x_* is the minimizer of v. Plugging x_* in the inequality above, we obtain $|\nabla v(0)||x| \geqslant \underline{\eta}|x|^2$. Thus, in view of Assumption (ii), we have

$$|x_*| \leqslant \frac{C_1}{\underline{\eta}}. \tag{E.4}$$

Denote $\tilde{v}(x) = v(x) - v(x_*)$. We have by definition $q = \exp(-\tilde{v})/\int \exp(-\tilde{v})$ and $\tilde{v}(x_*) = 0$ as well as $\nabla \tilde{v}(x_*) = 0$. It follows from Taylor expansion that

$$\frac{1}{2}\underline{\eta}|x - x_*|^2 \leqslant \tilde{v}(x) \leqslant \frac{1}{2}\bar{\eta}|x - x_*|^2,$$

so that

$$\left(\frac{\underline{\eta}}{2\pi}\right)^{d/2} \exp\left(-\frac{\bar{\eta}}{2}|x - x_*|^2\right) \leqslant q \leqslant \left(\frac{\bar{\eta}}{2\pi}\right)^{d/2} \exp\left(-\frac{\underline{\eta}}{2}|x - x_*|^2\right). \tag{E.5}$$

Step 2. Now we estimate the function r. Denote $\tilde{w}(x) = w(x) - w(x_*)$. We have by definition $r = \exp(-\tilde{w})/\int \exp(-\tilde{w})q$ and $\tilde{w}(x_*) = 0$. Thanks to Assumption (iii), we know that $\nabla w = \nabla \tilde{w}$ is uniformly bounded by C_2. Therefore it holds

$$-C_2|x - x_*| \leqslant \tilde{w}(x) \leqslant C_2|x - x_*|.$$

In particular, in view of (E.5) and (E.4), it holds for some $\underline{C}$, $\bar{C} > 0$,

$$\underline{C} \exp(-L|x - x_*|) \leqslant r \leqslant \bar{C} \exp(L|x - x_*|). \tag{E.6}$$

Step 3. Since $p = qr$, the conclusion follows immediately from (E.4), (E.5) and (E.6). □

E.3 Reflection coupling

In the section we recall the reflection coupling technique developped in [83, 85] and use it to estimate the $\mathcal{W}_1$–distance between the marginal laws of two diffusion processes with drift b and $b + \delta b$.

Assumption E.5. The drifts b and δb satisfy

(i) b and δb are Lipschitz in x, i.e., there is a constant $L > 0$ such that

$$|b(t,x)-b(t,y)|+|\delta b(t,x)-\delta b(t,y)| \leqslant L|x-y|, \qquad \text{for all } t \in [0,T],\ x,\ y \in \mathbb{R}^d;$$

(ii) there exists a continuous function $\kappa : (0,\infty) \to \mathbb{R}$ satisfying

$$\limsup_{r\to\infty} \kappa(r) < 0, \qquad \int_0^1 r\kappa^+(r)\,\mathrm{d}r < \infty$$

and

$$(x-y)\cdot\big(b(t,x)-b(t,y)\big) \leqslant \kappa(|x-y|)|x-y|^2, \qquad \text{for all } t \in [0,T],\ x,\ y \in \mathbb{R}^d.$$

Remark E.6. If $b(t,x) = -\big(\alpha(t,x) + \nabla\beta(t,x)\big)$ with α bounded and β η–convex in x, i.e.,

$$\big(\nabla\beta(t,x) - \nabla\beta(t,y)\big)\cdot(x-y) \geqslant \eta|x-y|^2,$$

then the function b satisfies Assumption E.5 (ii) with $\kappa(r) = 2\|\alpha\|_\infty r^{-1} - \eta$.

Theorem E.7. *Let Assumption E.5 hold. Consider the following two diffusion processes*

$$\mathrm{d}X_t = b(t,X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad \mathrm{d}Y_t = (b+\delta b)(t,Y_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t,$$

and denote their marginal distributions by $p^X_t := \mathcal{L}(X_t)$ and $p^Y_t := \mathcal{L}(Y_t)$. Then we have

$$\mathcal{W}_1(p^X_t, p^Y_t) \leqslant Ce^{-c\sigma^2 t}\left(\mathcal{W}_1(p^X_0, p^Y_0) + \int_0^t e^{c\sigma^2 s}\mathbb{E}\big[|\delta b(s,Y_s)|\big]\,\mathrm{d}s\right), \qquad \text{for all } t \geqslant 0, \tag{E.7}$$

where the constants C and c only depend on the function $\kappa(\cdot)/\sigma^2$.

Remark E.8. It follows immediately from Theorem E.7 that if π is an invariant distribution of the process X then

$$\mathcal{W}_1(p^X_t, \pi) \leqslant Ce^{-c\sigma^2 t}\mathcal{W}_1(p^X_0, \pi), \qquad \text{for all } t \geqslant 0.$$

In particular, it is unique and it is the limiting distribution of X.

Proof. We first recall the reflection-synchronuous coupling introduced in [85]. Introduce Lipschitz functions $\mathrm{rc} : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ and $\mathrm{sc} : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ satisfying

$$\mathrm{rc}^2(x,y) + \mathrm{sc}^2(x,y) = 1.$$

Fix a small constant $\eta > 0$. We impose that $\mathrm{rc}(x, y) = 1$ whenever $|x - y| > \eta$ and $\mathrm{rc}(x, y) = 0$ if $|x - y| \leqslant \eta/2$. The so-called reflection-synchronuous coupling is the strong solution to the following SDE system:

$$\begin{aligned}
\mathrm{d}X_t &= b(t, X_t)\,\mathrm{d}t + \mathrm{rc}(X_t, Y_t)\sigma\,\mathrm{d}W_t^1 + \mathrm{sc}(X_t, Y_t)\sigma\,\mathrm{d}W_t^2,\\
\mathrm{d}Y_t &= (b + \delta b)(t, Y_t)\,\mathrm{d}t + \mathrm{rc}(X_t, Y_t)(I - 2e_t\langle e_t, \cdot\rangle)\sigma\,\mathrm{d}W_t^1 + \mathrm{sc}(X_t, Y_t)\sigma\,\mathrm{d}W_t^2,
\end{aligned}$$

where W^1, W^2 are d-dimensional independent standard Brownian motion and

$$e_t = \frac{X_t - Y_t}{|X_t - Y_t|} \text{ for } X_t \neq Y_t \quad \text{and} \quad e_t = u \text{ for } X_t = Y_t,$$

with $u \in \mathbb{R}^d$ a fixed arbitrary unit vector. We denote by $\mathrm{rc}_t \coloneqq \mathrm{rc}(X_t, Y_t)$ and define $r_t \coloneqq |X_t - Y_t|$. Observe that

$$\mathrm{d}r_t = \langle e_t, b(t, X_t) - b(t, Y_t) - \delta b(t, Y_t)\rangle\,\mathrm{d}t + 2\,\mathrm{rc}_t\,\sigma\,\mathrm{d}W_t^\circ,$$

where W° is a one-dimensional standard Brownian motion, see [83, Lemma 6.2].

Next we construct an important auxiliary function f as in [85, Section 5.3]. First define two constants:

$$\begin{aligned}
R_1 &= \inf\{R \geqslant 0 \mid \kappa(r) \leqslant 0, \text{ for all } r \geqslant R\},\\
R_2 &= \inf\{R \geqslant R_1 \mid \kappa(r)R(R - R_1) \leqslant -4\sigma^2, \text{ for all } r \geqslant R\}.
\end{aligned}$$

Further define

$$\begin{aligned}
\varphi(r) &= \exp\biggl(-\frac{1}{2\sigma^2}\int_0^r u\kappa^+(u)\,\mathrm{d}u\biggr),\\
\Phi(r) &= \int_0^r \varphi(u)\,\mathrm{d}u,\\
g(r) &= 1 - \frac{c}{2}\int_0^r \Phi(u)\varphi(u)^{-1}\,\mathrm{d}u,
\end{aligned}$$

where the constant $c = \bigl(\int_0^{R_2}\Phi(r)\varphi(r)^{-1}\,\mathrm{d}r\bigr)^{-1}$, and eventually define the auxiliary function

$$f(r) = \int_0^r \varphi(u)g(u \wedge R_2)\,\mathrm{d}u.$$

One easily checks that

$$r\varphi(R_1) \leqslant \Phi(r) \leqslant 2f(r) \leqslant 2\Phi(r) \leqslant 2r, \qquad \text{for all } r > 0.$$

Note also that f is increasing and concave. In addition, f is linear on $[R_2, +\infty)$, twice continuously differentiable on $(0, R_2)$ and satisfies

$$2\sigma^2 f''(r) \leqslant -r\kappa^+(r)f'(r) - c\sigma^2 f(r), \qquad \text{for all } r \in (0, \infty) \setminus \{R_2\}. \tag{E.8}$$

This inequality follows easily by direct computation on $[0, R_2)$ and we refer to [85, Eqn (5.32)] for a detailed justification on $(R_2, +\infty)$. Then we have by the Itô–Tanaka formula as in [85, Eqn (5.26)] that

$$\begin{aligned}
\mathrm{d}f(r_t) \leqslant \bigl(f'_-(r_t)\langle e_t, b(t, X_t) - b(t, Y_t) - \delta b(t, Y_t)\rangle + 2\sigma^2\,\mathrm{rc}_t^2\, f''(r_t)\bigr)\,\mathrm{d}t&\\
+ 2\,\mathrm{rc}_t\, f'_-(r_t)\sigma\,\mathrm{d}W_t^\circ&.
\end{aligned}$$

Further note that

$$\langle e_t, b(t, X_t) - b(t, Y_t) \rangle \leqslant \mathbb{1}_{r_t < \eta} |b|_{\mathrm{Lip}} \eta + \mathbb{1}_{r_t \geqslant \eta} r_t \kappa^+(r_t).$$

Together with the fact that $f' \leqslant 1$, $f'' \leqslant 0$ and $\mathrm{rc}_t \, \mathbb{1}_{r_t \geqslant \eta} = 1$, we deduce that

$$\begin{aligned}
\mathrm{d} e^{c\sigma^2 t} f(r_t) \leqslant e^{c\sigma^2 t} \Big(& 2 \, \mathrm{rc}_t \, f'_-(r_t) \sigma \, \mathrm{d}W_t^\circ + |\delta b(t, Y_t)| \, \mathrm{d}t \\
& + \mathbb{1}_{r_t < \eta} (c\sigma^2 f(r_t) + |b|_{\mathrm{Lip}} \eta) \, \mathrm{d}t \\
& + \mathbb{1}_{r_t \geqslant \eta} \big(c\sigma^2 f(r_t) + r_t \kappa^+(r_t) f'(r_t) + 2\sigma^2 f''(r_t) \big) \, \mathrm{d}t \Big).
\end{aligned}$$

It follows from (E.8) that

$$\mathrm{d} e^{c\sigma^2 t} f(r_t) \leqslant e^{c\sigma^2 t} \Big(2 \, \mathrm{rc}_t \, f'_-(r_t) \sigma \, \mathrm{d}W_t^\circ + \big(|\delta b(t, Y_t)| + (c\sigma^2 + |b|_{\mathrm{Lip}}) \eta \big) \, \mathrm{d}t \Big).$$

Taking expectation on both sides, we obtain

$$\mathbb{E}[e^{c\sigma^2 t} f(r_t) - f(r_0)] \leqslant \int_0^t e^{c\sigma^2 s} \big(\mathbb{E}\big[|\delta b(s, Y_s)| \big] + (c\sigma^2 + |b|_{\mathrm{Lip}}) \eta \big) \, \mathrm{d}s.$$

Again due to the construction of f we have

$$\begin{aligned}
\mathcal{W}_1(p_t^X, p_t^Y) \leqslant \mathbb{E}[r_t] &\leqslant 2\varphi(R_1)^{-1} \mathbb{E}[f(r_t)] \\
&\leqslant 2\varphi(R_1)^{-1} e^{-c\sigma^2 t} \bigg(\mathbb{E}[f(r_0)] + \int_0^t e^{c\sigma^2 s} \mathbb{E}\big[|\delta b(s, Y_s)| \big] \, \mathrm{d}s \bigg) \\
&\qquad + 2\varphi(R_1)^{-1} \int_0^t e^{-c\sigma^2 (t-s)} (c\sigma^2 + |b|_{\mathrm{Lip}}) \eta \, \mathrm{d}s \\
&\leqslant 2\varphi(R_1)^{-1} e^{-c\sigma^2 t} \bigg(\mathcal{W}_1(p_0^X, p_0^Y) + \int_0^t e^{c\sigma^2 s} \mathbb{E}\big[|\delta b(s, Y_s)| \big] \, \mathrm{d}s \bigg) \\
&\qquad + 2\varphi(R_1)^{-1} \int_0^t e^{-c\sigma^2 (t-s)} (c\sigma^2 + |b|_{\mathrm{Lip}}) \eta \, \mathrm{d}s.
\end{aligned}$$

By passing to the limit $\eta \to 0$, we finally obtain the estimate (E.7). $\square$

Bibliography

[1] Shigeki Aida and Ichiro Shigekawa. Logarithmic Sobolev inequalities and spectral gaps: Perturbation theory. *J. Funct. Anal.*, 126(2):448–475, 1994. ISSN: 0022-1236. DOI: 10.1006/jfan.1994.1154.

[2] Zeyuan Allen-Zhu, Yuanzhi Li and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Adv. Neural Inf. Process. Syst. 32*, volume 32, pages 6158–6169. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/62dad6e273d32235ae02b7d321578ee8-Paper.pdf.

[3] Luigi Ambrosio, Nicola Fusco and Diego Pallara. *Functions of bounded variation and free discontinuity problems.* Oxford Math. Monogr. Clarendon Press, Oxford, 2000. ISBN: 0-19-850245-1.

[4] Luigi Ambrosio, Nicola Gigli and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Birkhäuser, Basel, 2nd edition, 2008. ISBN: 978-3-7643-8721-1.

[5] Christophe Andrieu, Nando de Freitas, Arnaud Doucet and Michael I. Jordan. An introduction to MCMC for machine learning. *Mach. Learn.*, 50(1-2):5–43, 2003. ISSN: 0885-6125. DOI: 10.1023/A:1020281327116.

[6] Cécile Ané, Sébastien Blachère, Djalil Chafaï, Pierre Fougères, Ivan Gentil, Florent Malrieu, Cyril Roberto and Grégory Scheffer. *Sur les inégalités de Sobolev logarithmiques.* French, volume 10 of *Panor. Synth.* Société mathématique de France, Paris, 2000. ISBN: 2-85629-105-8.

[7] Dyego Araújo, Roberto I. Oliveira and Daniel Yukimura. A mean-field limit for certain deep neural networks. 2019. arXiv: 1906.00193 [math.ST].

[8] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.*, 34(6):26–38, 2017. DOI: 10.1109/MSP.2017.2743240.

[9] Dominique Bakry, Franck Barthe, Patrick Cattiaux and Arnaud Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electron. Commun. Probab.*, 13:60–66, 2008. ISSN: 1083-589X. DOI: 10.1214/ECP.v13-1352.

[10] Dominique Bakry, Patrick Cattiaux and Arnaud Guillin. Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *J. Funct. Anal.*, 254(3):727–759, 2008. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2007.11.002.

[11] Dominique Bakry and Michel Émery. Diffusions hypercontractives. French. In *Sémin. de Probab. XIX*. Volume 1123, Lect. Notes Math. Springer-Verlag, Berlin etc., 1985.

[12] Dominique Bakry, Ivan Gentil and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren Math. Wiss.* Springer, Cham, 2014. ISBN: 978-3-319-00226-2; 978-3-319-00227-9. DOI: 10.1007/978-3-319-00227-9.

[13] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993. ISSN: 0018-9448. DOI: 10.1109/18.256500.

[14] Roland Bauerschmidt, Thierry Bodineau and Benoît Dagallier. A criterion on the free energy for log-Sobolev inequalities in mean-field particle systems. 2025. arXiv: 2503.24372 [math.PR].

[15] Michel Benaïm, Michel Ledoux and Olivier Raimond. Self-interacting diffusions. *Probab. Theory Relat. Fields*, 122(1):1–41, 2002. ISSN: 0178-8051. DOI: 10.1007/s004400100161.

[16] Michel Benaïm and Olivier Raimond. Self-interacting diffusions. II: Convergence in law. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 39(6):1043–1055, 2003. ISSN: 0246-0203.

[17] Michel Benaïm and Olivier Raimond. Self-interacting diffusions. III: Symmetric interactions. *Ann. Probab.*, 33(5):1716–1759, 2005. ISSN: 0091-1798. DOI: 10.1214/009117905000000251.

[18] Michel Benaïm and Olivier Raimond. Self-interacting diffusions. IV: Rate of convergence. *Electron. J. Probab.*, 16, 66, 2011. ISSN: 1083-6489. DOI: 10.1214/EJP.v16-948.

[19] Dario Benedetto, Emanuele Caglioti, José A. Carrillo and Mario Pulvirenti. A non-Maxwellian steady distribution for one-dimensional granular media. *J. Stat. Phys.*, 91(5-6):979–990, 1998. ISSN: 0022-4715. DOI: 10.1023/A:1023032000560.

[20] Armand Bernou and Mitia Duerinckx. Uniform-in-time estimates on the size of chaos for interacting brownian particles. 2024. arXiv: 2405.19306 [math.AP]. URL: https://arxiv.org/abs/2405.19306.

[21] George D. Birkhoff. Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. USA*, 17:656–660, 1931. ISSN: 0027-8424. DOI: 10.1073/pnas.17.12.656.

[22] Vladimir I. Bogachev, Nicolai V. Krylov, Michael Röckner and Stanislav V. Shaposhnikov. *Fokker–Planck–Kolmogorov equations*, volume 207 of *Math. Surv. Monogr.* American Mathematical Society, Providence, RI, 2015. ISBN: 978-1-4704-2558-6. DOI: 10.1090/surv/207.

[23] François Bolley, Ivan Gentil and Arnaud Guillin. Uniform convergence to equilibrium for granular media. *Arch. Ration. Mech. Anal.*, 208(2):429–445, 2013. ISSN: 0003-9527. DOI: 10.1007/s00205-012-0599-z.

[24] François Bolley, Arnaud Guillin and Florent Malrieu. Trend to equilibrium and particle approximation for a weakly selfconsistent Vlasov–Fokker–Planck equation. *ESAIM, Math. Model. Numer. Anal.*, 44(5):867–884, 2010. ISSN: 0764-583X. DOI: 10.1051/m2an/2010045.

[25] François Bolley and Cédric Villani. Weighted Csiszár–Kullback–Pinsker inequalities and applications to transportation inequalities. *Ann. Fac. Sci. Toulouse, Math. (6)*, 14(3):331–352, 2005. ISSN: 0240-2963. DOI: 10.5802/afst.1095.

[26] Nawaf Bou-Rabee and Andreas Eberle. Mixing time guarantees for unadjusted Hamiltonian Monte Carlo. *Bernoulli*, 29(1):75–104, 2023. ISSN: 1350-7265. DOI: 10.3150/21-BEJ1450.

[27] Stéphane Boucheron, Gábor Lugosi and Pascal Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013. ISBN: 978-0-19-953525-5.

[28] Jean Bourgain and Haïm Brezis. On the equation $\operatorname{div} Y = f$ and application to control of phases. *J. Am. Math. Soc.*, 16(2):393–426, 2003. ISSN: 0894-0347. DOI: 10.1090/S0894-0347-02-00411-3.

[29] Didier Bresch, Mitia Duerinckx and Pierre-Emmanuel Jabin. A duality method for mean-field limits with singular interactions, 2024. arXiv: 2402.04695 [math.AP].

[30] Didier Bresch, Pierre-Emmanuel Jabin and Juan Soler. A new approach to the mean-field limit of Vlasov–Fokker–Planck equations. 2023. arXiv: 2203.15747 [math.AP].

[31] Didier Bresch, Pierre-Emmanuel Jabin and Zhenfu Wang. Mean field limit and quantitative estimates with singular attractive kernels. *Duke Math. J.*, 172(13):2591–2641, 2023. ISSN: 0012-7094. DOI: 10.1215/00127094-2022-0088.

[32] Didier Bresch, Pierre-Emmanuel Jabin and Zhenfu Wang. Modulated free energy and mean field limit. In *Sémin. Laurent Schwartz, EDP Appl., Année 2019-2020*, Exposé n° II. DOI: 10.5802/slsedp.135.

[33] Didier Bresch, Pierre-Emmanuel Jabin and Zhenfu Wang. On mean-field limits and quantitative estimates with a large class of singular kernels: Application to the Patlak–Keller–Segel model. *C. R., Math., Acad. Sci. Paris*, 357(9):708–720, 2019. ISSN: 1631-073X. DOI: 10.1016/j.crma.2019.09.007.

[34] George W. Brown. Iterative solution of games by fictitious play. In Tjalling C. Koopmans, editor, *Activity Analysis of Production and Allocation*, chapter XXIV, pages 374–376. Chapman and Hall, London, 1951.

[35] Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry and Pierre-Louis Lions. *The master equation and the convergence problem in mean field games*, volume 201 of *Ann. Math. Stud.* Princeton University Press, Princeton, NJ, 2019. ISBN: 978-0-691-19070-9; 978-0-691-19071-6; 978-0-691-19371-7. DOI: 10.1515/9780691193717.

[36] Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: The fictitious play. *ESAIM, Control Optim. Calc. Var.*, 23(2):569–591, 2017. ISSN: 1292-8119. DOI: 10.1051/cocv/2016004.

[37] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications I. Mean field FBSDEs, control, and games*, volume 83 of *Probab. Theory Stoch. Model.* Springer, Cham, 2018. ISBN: 978-3-319-56437-1; 978-3-319-58920-6; 978-3-319-59820-8. DOI: 10.1007/978-3-319-58920-6.

[38] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications II. Mean field games with common noise and master equations*, volume 84 of *Probab. Theory Stoch. Model.* Springer, Cham, 2018. ISBN: 978-3-319-56435-7; 978-3-319-56436-4; 978-3-319-59820-8. DOI: 10.1007/978-3-319-56436-4.

[39] José A. Carrillo, Robert J. McCann and Cédric Villani. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoam.*, 19(3):971–1018, 2003. ISSN: 0213-2230. DOI: 10.4171/RMI/376.

[40] Patrick Cattiaux and Arnaud Guillin. Functional inequalities for perturbed measures with applications to log-concave measures and to some Bayesian problems. *Bernoulli*, 28(4):2294–2321, 2022. ISSN: 1350-7265. DOI: 10.3150/21-BEJ1419.

[41] Patrick Cattiaux and Arnaud Guillin. Semi log-concave Markov diffusions. In *Séminaire de Probabilités XLVI*, pages 231–292. Springer, Cham, 2014. ISBN: 978-3-319-11969-4; 978-3-319-11970-0. DOI: 10.1007/978-3-319-11970-0_9.

[42] Patrick Cattiaux, Arnaud Guillin and Pierre-André Zitt. Poincaré inequalities and hitting times. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 49(1):95–118, 2013. ISSN: 0246-0203. DOI: 10.1214/11-AIHP447.

[43] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. I: Models and methods. *Kinet. Relat. Models*, 15(6):895–1015, 2022. ISSN: 1937-5093. DOI: 10.3934/krm.2022017.

[44] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. II: Applications. *Kinet. Relat. Models*, 15(6):1017–1173, 2022. ISSN: 1937-5093. DOI: 10.3934/krm.2022018.

[45] Nicolas Champagnat and Pierre-Emmanuel Jabin. Strong solutions to stochastic differential equations with rough coefficients. *Ann. Probab.*, 46(3):1498–1541, 2018. ISSN: 0091-1798. DOI: 10.1214/17-AOP1208.

[46] Paul-Éric Chaudru de Raynal, Manh Hong Duong, Pierre Monmarché, Milica Tomašević and Julian Tugaut. Reducing exit-times of diffusions with repulsive interactions. *ESAIM, Probab. Stat.*, 27:723–748, 2023. ISSN: 1292-8100. DOI: 10.1051/ps/2023012.

[47] Mu-Fa Chen and Feng-Yu Wang. Estimation of spectral gap for elliptic operators. *Trans. Am. Math. Soc.*, 349(3):1239–1267, 1997. ISSN: 0002-9947. DOI: 10.1090/S0002-9947-97-01812-6.

[48] Fan Chen, Yiqing Lin, Zhenjie Ren and Songbo Wang. Uniform-in-time propagation of chaos for kinetic mean field Langevin dynamics. *Electron. J. Probab.*, 29, 17, 2024. DOI: 10.1214/24-EJP1079.

[49] Fan Chen, Zhenjie Ren and Songbo Wang. Entropic fictitious play for mean field optimization problem. *J. Mach. Learn. Res.*, 24, 211, 2023. URL: http://jmlr.org/papers/v24/22-0411.html.

[50] Fan Chen, Zhenjie Ren and Songbo Wang. Uniform-in-time propagation of chaos for mean field Langevin dynamics. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 2024. arXiv: 2212.03050 [math.PR]. Forthcoming.

[51] Li Chen, Jinwook Jung, Peter Pickl and Zhenfu Wang. On the mean-field limit of Vlasov–Poisson–Fokker–Planck equations. 2025. arXiv: 2505.13038 [math-ph]. URL: https://arxiv.org/abs/2505.13038.

[52] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for Markov chains. 2022. arXiv: 2203.04163 [math.PR]. URL: https://arxiv.org/abs/2203.04163.

[53] Zixiang Chen, Yuan Cao, Quanquan Gu and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. In *Adv. Neural Inf. Process. Syst.* Volume 33, pages 13363–13373. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/9afe487de556e59e6db6c862adfe25a4-Paper.pdf.

[54] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 2018. URL: https://proceedings.mlr.press/v75/cheng18a.html.

[55] Sinho Chewi, Atsushi Nitanda and Matthew S. Zhang. Uniform-in-N log-Sobolev inequality for the mean-field Langevin dynamics with convex energy. 2024. arXiv: 2409.10440 [math.PR].

[56] Lénaïc Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. *Trans. Mach. Learn. Res.*, 2022. ISSN: 2835-8856. URL: https://openreview.net/forum?id=BDqzLH1gEm.

[57] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Adv. Neural Inf. Process. Syst.* Volume 31, pages 3036–3046. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

[58] Lénaïc Chizat, Edouard Oyallon and Francis Bach. On lazy training in differentiable programming. In *Adv. Neural Inf. Process. Syst.* Volume 32, pages 2937–2947. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

[59] Antonin Chodron de Courcel, Matthew Rosenzweig and Sylvia Serfaty. Sharp uniform-in-time mean-field convergence for singular periodic Riesz flows. *Ann. Inst. Henri Poincaré C, Anal. Non Linéaire*, 42(2):391–472, 2025. ISSN: 0294-1449. DOI: 10.4171/AIHPC/105.

[60] Julien Claisse, Giovanni Conforti, Zhenjie Ren and Songbo Wang. Mean field optimization problem regularized by Fisher information. 2023. arXiv: 2302.05938 [math.PR].

[61] Giovanni Conforti. Coupling by reflection for controlled diffusion processes: Turnpike property and large time behavior of Hamilton–Jacobi–Bellman equations. *Ann. Appl. Probab.*, 33(6A):4608–4644, 2023. ISSN: 1050-5164. DOI: 10.1214/22-AAP1927.

[62] Giovanni Conforti. Weak semiconvexity estimates for Schrödinger potentials and logarithmic Sobolev inequality for Schrödinger bridges. *Probab. Theory Relat. Fields*, February 2024. ISSN: 0178-8051, 1432-2064. DOI: 10.1007/s00440-024-01264-6.

[63] Giovanni Conforti, Anna Kazeykina and Zhenjie Ren. Game on random environment, mean-field Langevin system, and neural networks. *Math. Oper. Res.*, 48(1):78–99, 2023. DOI: 10.1287/moor.2022.1252.

[64] Michael G. Crandall, Hitoshi Ishii and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. *Bull. Am. Math. Soc., New Ser.*, 27(1):1–67, 1992. ISSN: 0273-0979. DOI: 10.1090/S0273-0979-1992-00266-5.

[65] Michael Cranston. Gradient estimates on manifolds using coupling. *J. Funct. Anal.*, 99(1):110–124, 1991. ISSN: 0022-1236. DOI: 10.1016/0022-1236(91)90054-9.

[66] Michael Cranston and Yves Le Jan. Self attracting diffusions: Two case studies. *Math. Ann.*, 303(1):87–93, 1995. ISSN: 0025-5831. DOI: 10.1007/BF01460980.

[67] Noé Cuneo and Jean-Pierre Eckmann. Non-equilibrium steady states for chains of four rotors. *Commun. Math. Phys.*, 345(1):185–221, 2016. ISSN: 0010-3616. DOI: 10.1007/s00220-015-2550-2.

[68] George Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989. ISSN: 0932-4194. DOI: 10.1007/BF02551274.

[69] Antonin Chodron de Courcel, Matthew Rosenzweig and Sylvia Serfaty. The attractive log gas: stability, uniqueness, and propagation of chaos, 2023. arXiv: 2311.14560 [math.AP].

[70] François Delarue and Alvin Tse. Uniform in time weak propagation of chaos on the torus. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 61(2):1021–1074, 2025. ISSN: 0246-0203,1778-7017. DOI: 10.1214/23-aihp1451.

[71] Luc Devroye, Abbas Mehrabian and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. 2023. arXiv: 1810.08693 [math.ST].

[72] Joscha Diehl, Peter K. Friz and Harald Oberhauser. Parabolic comparison revisited and applications. 2011. arXiv: 1102.5774 [math.AP].

[73] Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *J. Am. Stat. Assoc.*, 116(535):1507–1520, 2021. ISSN: 0162-1459. DOI: 10.1080/01621459.2020.1745812.

[74] Jérôme Droniou and Robert Eymard. Uniform-in-time convergence of numerical methods for non-linear degenerate parabolic equations. *Numer. Math.*, 132(4):721–766, 2016. ISSN: 0029-599X. DOI: 10.1007/s00211-015-0733-6.

[75] Kai Du, Yifan Jiang and Jinfeng Li. Empirical approximation to invariant measures for McKean–Vlasov processes: Mean-field interaction vs self-interaction. *Bernoulli*, 29(3):2492–2518, 2023. ISSN: 1350-7265. DOI: 10.3150/22-BEJ1550.

[76] Kai Du, Yifan Jiang and Xiaochen Li. Sequential propagation of chaos. 2023. arXiv: 2301.09913 [math.PR].

[77] Kai Du, Zhenjie Ren, Florin Suciu and Songbo Wang. Self-interacting approximation to McKean–Vlasov long-time limit: a Markov chain Monte Carlo method. 2023. arXiv: 2311.11428 [math.PR].

[78] Mitia Duerinckx. On the size of chaos via Glauber calculus in the classical mean-field dynamics. *Commun. Math. Phys.*, 382(1):613–653, 2021. ISSN: 0010-3616. DOI: 10.1007/s00220-021-03978-3.

[79] Alain Durmus, Andreas Eberle, Arnaud Guillin and Katharina Schuh. Sticky nonlinear SDEs and convergence of McKean–Vlasov equations without confinement. *Stoch. PDE Anal. Comp.*, November 2023. ISSN: 2194-0401, 2194-041X. DOI: 10.1007/s40072-023-00315-8.

[80] Alain Durmus, Andreas Eberle, Arnaud Guillin and Raphael Zimmer. An elementary approach to uniform in time propagation of chaos. *Proc. Am. Math. Soc.*, 148(12):5387–5398, 2020. ISSN: 0002-9939. DOI: 10.1090/proc/14612.

[81] Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019. ISSN: 1350-7265. DOI: 10.3150/18-BEJ1073.

[82] Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017. ISSN: 1050-5164. DOI: 10.1214/16-AAP1238.

[83] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Relat. Fields*, 166(3-4):851–886, 2016. ISSN: 0178-8051. DOI: 10.1007/s00440-015-0673-1.

[84] Andreas Eberle, Arnaud Guillin and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.*, 47(4):1982–2010, 2019. ISSN: 0091-1798. DOI: 10.1214/18-AOP1299.

[85] Andreas Eberle, Arnaud Guillin and Raphael Zimmer. Quantitative Harris-type theorems for diffusions and McKean-Vlasov processes. *Trans. Am. Math. Soc.*, 371(10):7135–7173, 2019. ISSN: 0002-9947. DOI: 10.1090/tran/7576.

[86] Andreas Eberle and Raphael Zimmer. Sticky couplings of multidimensional diffusions with different drifts. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 55(4):2370–2394, 2019. ISSN: 0246-0203. DOI: 10.1214/18-AIHP951.

[87] Eberhard Engel and Reiner M. Dreizler. *Density functional theory. An advanced course.* Theor. Math. Phys. Springer, Berlin, 2011. ISBN: 978-3-642-14089-1; 978-3-642-14090-7. DOI: 10.1007/978-3-642-14090-7.

[88] Helmut Eschrig. *The fundamentals of density functional theory*, number 1. Vieweg+Teubner Verlag Wiesbaden, 1996.

[89] Cong Fang, Jason Lee, Pengkun Yang and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1887–1936. PMLR, 2021. URL: https://proceedings.mlr.press/v134/fang21a.html.

[90] Xuanrui Feng and Zhenfu Wang. Quantitative propagation of chaos for 2D viscous vortex model on the whole space. 2023. arXiv: 2310.05156 [math.AP].

[91] Wendell H. Fleming and H. Mete Soner. *Controlled Markov processes and viscosity solutions*, volume 25 of *Stoch. Model. Appl. Probab.* Springer, New York, NY, 2nd edition, 2006. ISBN: 0-387-26045-5; 978-1-4419-2078-2; 0-387-31071-1. DOI: 10.1007/0-387-31071-1.

[92] Wendell H. Fleming and Michel Viot. Some measure-valued Markov processes in population genetics theory. *Indiana Univ. Math. J.*, 28:817–843, 1979. ISSN: 0022-2518. DOI: 10.1512/iumj.1979.28.28058.

[93] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Relat. Fields*, 162(3-4):707–738, 2015. ISSN: 0178-8051. DOI: 10.1007/s00440-014-0583-7.

[94] Wilfrid Gangbo and Alpár Richárd Mészáros. Global well-posedness of master equations for deterministic displacement convex potential mean field games. *Commun. Pure Appl. Math.*, 75(12):2685–2801, 2022. ISSN: 0010-3640. DOI: 10.1002/cpa.22069.

[95] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Adv. Neural Inf. Process. Syst.* Volume 32, pages 9111–9121. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/c133fb1bb634af68c5088f3438848bfd-Paper.pdf.

[96] Yoav Goldberg. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.*, 57:345–420, 2016. ISSN: 1076-9757. DOI: 10.1613/jair.4992.

[97] Arnaud Guillin, Pierre Le Bris and Pierre Monmarché. Convergence rates for the Vlasov–Fokker–Planck equation and uniform in time propagation of chaos in non convex cases. *Electron. J. Probab.*, 27, 124, 2022. ISSN: 1083-6489. DOI: 10.1214/22-EJP853.

[98] Arnaud Guillin, Pierre Le Bris and Pierre Monmarché. Uniform in time propagation of chaos for the 2d vortex model and other singular stochastic systems. *J. Eur. Math. Soc.*, 27(6):2359–2386, 2025. ISSN: 1435-9855. DOI: 10.4171/JEMS/1413.

[99] Arnaud Guillin, Wei Liu, Liming Wu and Chaoen Zhang. The kinetic Fokker–Planck equation with mean field interaction. *J. Math. Pures Appl. (9)*, 150:1–23, 2021. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2021.04.001.

[100] Arnaud Guillin, Wei Liu, Liming Wu and Chaoen Zhang. Uniform Poincaré and logarithmic Sobolev inequalities for mean field particle systems. *Ann. Appl. Probab.*, 32(3):1590–1614, 2022. ISSN: 1050-5164. DOI: 10.1214/21-AAP1707.

[101] Arnaud Guillin and Pierre Monmarché. Uniform long-time and propagation of chaos estimates for mean field kinetic particles in non-convex landscapes. *J. Stat. Phys.*, 185(2), 15, 2021. ISSN: 0022-4715. DOI: 10.1007/s10955-021-02839-6.

[102] Arnaud Guillin and Feng-Yu Wang. Degenerate Fokker-Planck equations: Bismut formula, gradient estimate and Harnack inequality. *J. Differ. Equations*, 253(1):20–40, 2012. ISSN: 0022-0396. DOI: 10.1016/j.jde.2012.03.014.

[103] Yan Guo. The Landau equation in a periodic box. *Commun. Math. Phys.*, 231(3):391–434, 2002. ISSN: 0010-3616. DOI: 10.1007/s00220-002-0729-9.

[104] Saeed Hadikhanloo and Francisco J. Silva. Finite mean field games: Fictitious play and convergence to a first order continuous mean field game. *J. Math. Pures Appl. (9)*, 132:369–397, 2019. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2019.02.006.

[105] Yi Han. Entropic propagation of chaos for mean field diffusion with L^p interactions via hierarchy, linear growth and fractional noise. 2023. arXiv: 2205.02772 [math.PR].

[106] Trevor Hastie, Andrea Montanari, Saharon Rosset and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.*, 50(2):949–986, 2022. ISSN: 0090-5364. DOI: 10.1214/21-AOS2133.

[107] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conf. Comput. Vis. Pattern Recognit.* Pages 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[108] Dan Henry. *Geometric theory of semilinear parabolic equations*, volume 840 of *Lect. Notes Math.* Springer, Cham, 1981.

[109] Frédéric Hérau. Short and long time behavior of the Fokker-Planck equation in a confining potential and applications. *J. Funct. Anal.*, 244(1):95–118, 2007. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2006.11.013.

[110] Samuel Herrmann and Julian Tugaut. Non-uniqueness of stationary measures for self-stabilizing processes. *Stochastic Process. Appl.*, 120(7):1215–1246, 2010. ISSN: 0304-4149. DOI: 10.1016/j.spa.2010.03.009.

[111] Elias Hess-Childs and Keefer Rowan. Higher-order propagation of chaos in L^2 for interacting diffusions. *Probab. Math. Phys.*, 6(2):581–646, 2025. ISSN: 2690-0998. DOI: 10.2140/pmp.2025.6.581.

[112] Geoffrey Hinton, Nitish Srivastava and Kevin Swersky. Neural networks for machine learning: Lecture 6a: Overview of mini-batch gradient descent. 2012. URL: http://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf (visited on 06/23/2023).

[113] Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *J. Stat. Phys.*, 46(5-6):1159–1194, 1987. ISSN: 0022-4715. DOI: 10.1007/BF01011161.

[114] Lars Hörmander. *The analysis of linear partial differential operators. I. Distribution theory and Fourier analysis.* Volume 256 of *Grundlehren Math. Wiss.* Springer-Verlag, Berlin etc., 2nd edition, 1990. ISBN: 3-540-52343-X; 3-540-52345-6.

[115] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[116] Kurt Hornik, Maxwell Stinchcombe and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, 1989. ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8.

[117] Kaitong Hu, Zhenjie Ren, David Šiška and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 57(4):2043–2065, 2021. ISSN: 0246-0203. DOI: 10.1214/20-AIHP1140.

[118] Gao Huang, Zhuang Liu, Laurens van der Maaten and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[119] Jiaoyang Huang, Matthew Rosenzweig and Sylvia Serfaty. The modulated free energy method on $\mathbb{R}^d$. In preparation.

[120] Xing Huang. Coupling by change of measure for conditional McKean-Vlasov SDEs and applications. *Stochastic Processes Appl.*, 179:17, 104508, 2025. ISSN: 0304-4149. DOI: 10.1016/j.spa.2024.104508.

[121] Xing Huang, Eva Kopfer and Panpan Ren. Log-Sobolev inequality for decoupled and McKean–Vlasov SDEs and application on exponential ergodicity. 2025. arXiv: 2501.16092 [math.PR]. URL: https://arxiv.org/abs/2501.16092.

[122] Alessandra Iacobucci, Stefano Olla and Gabriel Stoltz. Convergence rates for nonequilibrium Langevin dynamics. *Ann. Math. Qué.*, 43(1):73–98, 2019. ISSN: 2195-4755. DOI: 10.1007/s40316-017-0091-0.

[123] Alessandra Iacobucci, Stefano Olla and Gabriel Stoltz. Thermo-mechanical transport in rotor chains. *J. Stat. Phys.*, 183(2), 26, 2021. ISSN: 0022-4715. DOI: 10.1007/s10955-021-02748-8.

[124] Pierre-Emmanuel Jabin and Zhenfu Wang. Quantitative estimates of propagation of chaos for stochastic systems with $W^{-1,\infty}$ kernels. *Invent. Math.*, 214(1):523–591, 2018. ISSN: 0020-9910. DOI: 10.1007/s00222-018-0808-y.

[125] Arthur Jacot, Franck Gabriel and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Adv. Neural Inf. Process. Syst.* Volume 31, pages 8571–8580. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

[126] Richard Jordan, David Kinderlehrer and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998. ISSN: 0036-1410. DOI: 10.1137/S0036141096303359.

[127] Christos Karras, Aristeidis Karras, Markos Avlonitis and Spyros Sioutas. An overview of MCMC methods: From theory to applications. In *Artif. Intell. Appl. Innov. AIAI 2022 IFIP WG 12.5 Int. Workshops*, pages 319–332, Cham. Springer International Publishing, 2022. ISBN: 978-3-031-08341-9.

[128] Anna Kazeykina, Zhenjie Ren, Xiaolu Tan and Junjian Yang. Ergodicity of the underdamped mean-field Langevin dynamics. *Ann. Appl. Probab.*, 34(3):3181–3226, 2024. ISSN: 1050-5164. DOI: 10.1214/23-AAP2036.

[129] Evelyn F. Keller and Lee A. Segel. Model for chemotaxis. *J. Theor. Biol.*, 30(2):225–234, 1971. ISSN: 0022-5193. DOI: 10.1016/0022-5193(71)90050-6.

[130] Wilfrid S. Kendall. The radial part of Brownian motion on a manifold: A semimartingale property. *Ann. Probab.*, 15:1491–1500, 1987. ISSN: 0091-1798. DOI: 10.1214/aop/1176991988.

[131] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2017. arXiv: 1412.6980 [cs.LG].

[132] Victor Kleptsyn and Aline Kurtzmann. Ergodicity of self-attracting motion. *Electron. J. Probab.*, 17, 50, 2012. ISSN: 1083-6489. DOI: 10.1214/EJP.v17-2121.

[133] Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *J. Mach. Learn. Res.*, 22(17):1–40, 2021. URL: http://jmlr.org/papers/v22/19-466.html.

[134] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN: 0001-0782. DOI: 10.1145/3065386.

[135] Anders Krogh and John Hertz. A simple weight decay can improve generalization. In *Adv. Neural Inf. Process. Syst.* Volume 4, pages 950–957. Morgan-Kaufmann, 1991. URL: https://proceedings.neurips.cc/paper_files/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf.

[136] Nicolai V. Krylov and Michael Röckner. Strong solutions of stochastic equations with singular time dependent drift. *Probab. Theory Relat. Fields*, 131(2):154–196, 2005. ISSN: 0178-8051. DOI: 10.1007/s00440-004-0361-z.

[137] Alexei Kulik and Ilya Pavlyukevich. Non-Gaussian limit theorem for non-linear Langevin equations driven by Lévy noise. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 55(3):1278–1315, 2019. ISSN: 0246-0203. DOI: 10.1214/18-AIHP919.

[138] Yoshiki Kuramoto. Rhythms and turbulence in populations of chemical oscillators. *Phys. A, Stat. Mech. Appl.*, 106(1):128–143, 1981. ISSN: 0378-4371. DOI: 10.1016/0378-4371(81)90214-4.

[139] Kazumasa Kuwada. Duality on gradient estimates and Wasserstein controls. *J. Funct. Anal.*, 258(11):3758–3774, 2010. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2010.01.010.

[140] Daniel Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *Probab. Math. Phys.*, 4(2):377–432, 2023. ISSN: 2690-0998. DOI: 10.2140/pmp.2023.4.377.

[141] Daniel Lacker. Quantitative approximate independence for continuous mean field Gibbs measures. *Electron. J. Probab.*, 27, 15, 2022. ISSN: 1083-6489. DOI: 10.1214/22-EJP743.

[142] Daniel Lacker and Luc Le Flem. Sharp uniform-in-time propagation of chaos. *Probab. Theory Relat. Fields*, 187(1-2):443–480, 2023. ISSN: 0178-8051. DOI: 10.1007/s00440-023-01192-x.

[143] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Jpn. J. Math.*, 2(1):229–260, 2007. ISSN: 0289-2316. URL: https://doi.org/10.1007/s11537-007-0657-8.

[144] Pierre Lavigne and Laurent Pfeiffer. Generalized conditional gradient and learning in potential mean field games. *Appl. Math. Optim.*, 88(3), 89, 2023. ISSN: 0095-4616. DOI: 10.1007/s00245-023-10056-8.

[145] Pierre Le Bris. *Systems of particles in (singular) interaction : long time behavior and propagation of chaos.* PhD thesis, Sorbonne Université, June 2023. URL: https://theses.hal.science/tel-04357166.

[146] Yann LeCun, Léon Bottou, Yoshua Bengio and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. DOI: 10.1109/5.726791.

[147] Yann LeCun, Corinna Cortes and Christopher J. C. Burges. The MNIST database of handwritten digits. 1998. URL: http://yann.lecun.com/exdb/mnist/ (visited on 06/23/2023).

[148] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Math. Surv. Monogr.* American Mathematical Society, Providence, RI, 2001. ISBN: 0-8218-2864-9.

[149] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces. Isoperimetry and processes*, volume 23 of *Ergeb. Math. Grenzgeb., 3. Folge*. Springer-Verlag, Berlin etc., 1991. ISBN: 3-540-52013-9.

[150] Tony Lelièvre, Mathias Rousset and Gabriel Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21(6):1155–1181, 2008. ISSN: 0951-7715. DOI: 10.1088/0951-7715/21/6/001.

[151] Mathieu Lewin. *Théorie spectrale et mécanique quantique*. French, volume 87 of *Math. Appl.* Springer, Cham, 2022. ISBN: 978-3-030-93435-4; 978-3-030-93436-1. DOI: 10.1007/978-3-030-93436-1.

[152] Torgny Lindvall and L. Christopher G. Rogers. Coupling of multidimensional diffusions by reflection. *Ann. Probab.*, 14:860–872, 1986. ISSN: 0091-1798. DOI: 10.1214/aop/1176992442.

[153] Tianyi Liu, Yan Li, Song Wei, Enlu Zhou and Tuo Zhao. Noisy gradient descent converges to flat minima for nonconvex matrix factorization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1891–1899. PMLR, 2021.

[154] Yanli Liu, Yuan Gao and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Adv. Neural Inf. Process. Syst.* Volume 33, pages 18261–18271. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/d3f5d4de09ea19461dab00590df91e4f-Paper.pdf.

[155] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[156] Yulong Lu. Two-scale gradient descent ascent dynamics finds mixed Nash equilibria of continuous games: A mean-field perspective. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22790–22811. PMLR, 2023. URL: https://proceedings.mlr.press/v202/lu23b.html.

[157] Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021. ISSN: 1350-7265. DOI: 10.3150/20-BEJ1297.

[158] Jin Ma and Jiongmin Yong. *Forward-backward stochastic differential equations and their applications*, volume 1702 of *Lect. Notes Math.* Springer, Berlin, 1999. ISBN: 3-540-65960-9. DOI: 10.1007/978-3-540-48831-6.

[159] Florent Malrieu. Logarithmic Sobolev inequalities for some nonlinear PDE's. *Stochastic Processes Appl.*, 95(1):109–132, 2001. ISSN: 0304-4149. DOI: 10.1016/S0304-4149(01)00095-3.

[160] Fabio Marchesoni. A nonlinear langevin equation for fluid relaxation. *Prog. Theor. Phys.*, 74(6):1339–1342, 1985. ISSN: 0033-068X. DOI: 10.1143/PTP.74.1339.

[161] Nicolas Martzel and Claude Aslangul. Mean-field treatment of the many-body Fokker-Planck equation. *J. Phys. A, Math. Gen.*, 34(50):11225–11240, 2001. ISSN: 0305-4470. DOI: 10.1088/0305-4470/34/50/305.

[162] Song Mei, Theodor Misiakiewicz and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the 32rd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 2019. URL: https://proceedings.mlr.press/v99/mei19a.html.

[163] Song Mei, Andrea Montanari and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA*, 115(33):e7665–e7671, 2018. ISSN: 0027-8424. DOI: 10.1073/pnas.1806579115.

[164] Dov Monderer and Lloyd S. Shapley. Fictitious play property for games with identical interests. *J. Econ. Theory*, 68(1):258–265, 1996. ISSN: 0022-0531. DOI: 10.1006/jeth.1996.0014.

[165] Pierre Monmarché. Almost sure contraction for diffusions on $\mathbb{R}^d$. Application to generalized Langevin diffusions. *Stochastic Process. Appl.*, 161:316–349, 2023. ISSN: 0304-4149. DOI: 10.1016/j.spa.2023.04.006.

[166] Pierre Monmarché. Generalized Γ calculus and application to interacting particles on a graph. *Potential Anal.*, 50(3):439–466, 2019. ISSN: 0926-2601. DOI: 10.1007/s11118-018-9689-3.

[167] Pierre Monmarché. Long-time behaviour and propagation of chaos for mean field kinetic particles. *Stochastic Processes Appl.*, 127(6):1721–1737, 2017. ISSN: 0304-4149. DOI: 10.1016/j.spa.2016.10.003.

[168] Pierre Monmarché. Wasserstein contraction and Poincaré inequalities for elliptic diffusions with high diffusivity. *Ann. Henri Lebesgue*, 6:941–973, 2023. DOI: 10.5802/ahl.182.

[169] Pierre Monmarché and Mouad Ramil. Overdamped limit at stationarity for non-equilibrium Langevin diffusions. *Electron. Commun. Probab.*, 27, 3, 2022. ISSN: 1083-589X. DOI: 10.1214/22-ECP447.

[170] Pierre Monmarché, Zhenjie Ren and Songbo Wang. Time-uniform log-Sobolev inequalities and applications to propagation of chaos. *Electron. J. Probab.*, 29, 154, 2024. DOI: 10.1214/24-EJP1217.

[171] Pierre Monmarché and Songbo Wang. Logarithmic Sobolev inequalities for non-equilibrium steady states. *Potential Anal.*, 2025. DOI: 10.1007/s11118-025-10211-6. arXiv: 2309.07858 [math.PR]. Pre-published.

[172] Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*. Pages 113–162. CRC Press, Boca Raton, FL, 2011. ISBN: 978-1-4200-7941-8; 978-1-4200-7942-5.

[173] Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach and James Martens. Adding gradient noise improves learning for very deep networks. 2015. arXiv: 1511.06807 [stat.ML].

[174] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program., Ser. A*, 103:127–152, 2005. ISSN: 0025-5610. DOI: 10.1007/s10107-004-0552-5.

[175] Phan-Minh Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. 2019. arXiv: 1902.02880 [cs.LG].

[176] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *Math. Stat. Learn.*, 6(3/4):201–357, 2023. ISSN: 2520-2316. DOI: 10.4171/MSL/42.

[177] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. 2017. arXiv: 1712.05438 [stat.ML].

[178] Atsushi Nitanda, Denny Wu and Taiji Suzuki. Convex analysis of the mean field Langevin dynamics. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9741–9757. PMLR, 2022. URL: https://proceedings.mlr.press/v151/nitanda22a.html.

[179] Atsushi Nitanda, Denny Wu and Taiji Suzuki. Particle dual averaging: optimization of mean field neural network with global convergence rate analysis. *J. Stat. Mech. Theory Exp.*, 2022(11), 114010, 2022. ISSN: 1742-5468. DOI: 10.1088/1742-5468/ac98a8.

[180] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000. ISSN: 0022-1236. DOI: 10.1006/jfan.2000.3557.

[181] Giorgio Parisi. Correlation functions and computer simulations. *Nucl. Phys. B*, 180(3):378–384, 1981. ISSN: 0550-3213. DOI: 10.1016/0550-3213(81)90056-0.

[182] Thierry Paul, Mario Pulvirenti and Sergio Simonella. On the size of chaos in the mean field dynamics. *Arch. Ration. Mech. Anal.*, 231(1):285–317, 2019. ISSN: 0003-9527. DOI: 10.1007/s00205-018-1280-y.

[183] Sarah Perrin, Julien Perolat, Mathieu Laurière, Matthieu Geist, Romuald Élie and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. In *Adv. Neural Inf. Process. Syst.* Volume 33, pages 13199–13213. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/995ca733e3657ff9f5f3c823d73371e1-Paper.pdf.

[184] Huy Tuan Pham and Phan-Minh Nguyen. Global convergence of three-layer neural networks in the mean field regime. In *9th International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=KvyxFqZS_D.

[185] Boris T. Poljak. Some methods of speeding up the convergence of iteration methods. *U.S.S.R. Comput. Math. Math. Phys.*, 4(5):1–17, 1967. ISSN: 0041-5553. DOI: 10.1016/0041-5553(64)90137-5.

[186] Alessio Porretta and Enrico Priola. Global Lipschitz regularizing effects for linear and nonlinear parabolic equations. *J. Math. Pures Appl. (9)*, 100(5):633–686, 2013. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2013.01.016.

[187] Enrico Priola and Feng-Yu Wang. Gradient estimates for diffusion semigroups with singular coefficients. *J. Funct. Anal.*, 236(1):244–264, 2006. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2005.12.010.

[188] Olivier Raimond. Self-attracting diffusions: Case of the constant interaction. *Probab. Theory Relat. Fields*, 107(2):177–196, 1997. ISSN: 0178-8051. DOI: 10.1007/s004400050082.

[189] Sashank J. Reddi, Satyen Kale and Sanjiv Kumar. On the convergence of Adam and beyond. In *6th International Conference on Learning Representations*, 2018. URL: https://openreview.net/forum?id=ryQu7f-RZ.

[190] Michael Reed and Barry Simon. Methods of Modern Mathematical Physics. 4 volumes. Academic Press, 1980.

[191] Michael Reed and Barry Simon. *Methods of modern mathematical physics. II. Fourier analysis, self-adjointness.* Academic Press, New York etc., 1975, pages xv+361.

[192] Panpan Ren and Feng-Yu Wang. Entropy estimate between diffusion processes with application to nonlinear Fokker–Planck equations. 2023. arXiv: 2302.13500 [math.PR].

[193] Panpan Ren and Feng-Yu Wang. Exponential convergence in entropy and Wasserstein for McKean–Vlasov SDEs. *Nonlinear Anal., Theory Methods Appl., Ser. A, Theory Methods*, 206, 112259, 2021. ISSN: 0362-546X. DOI: 10.1016/j.na.2021.112259.

[194] Zhenjie Ren and Songbo Wang. Size of chaos for Gibbs measures of mean field interacting diffusions. 2024. arXiv: 2411.14236 [math.PR].

[195] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 60(1):255–268, 1998. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00123.

[196] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN: 1350-7265. DOI: 10.2307/3318418.

[197] Julia Robinson. An iterative method of solving a game. *Ann. Math. (2)*, 54:296–301, 1951. ISSN: 0003-486X. DOI: 10.2307/1969530.

[198] Michael Röckner and Feng-Yu Wang. Weak Poincaré inequalities and L^2-convergence rates of Markov semigroups. *J. Funct. Anal.*, 185(2):564–603, 2001. ISSN: 0022-1236. DOI: 10.1006/jfan.2001.3776.

[199] Matthew Rosenzweig and Sylvia Serfaty. Global-in-time mean-field convergence for singular Riesz-type diffusive flows. *Ann. Appl. Probab.*, 33(2):954–998, 2023. ISSN: 1050-5164. DOI: 10.1214/22-AAP1833.

[200] Matthew Rosenzweig and Sylvia Serfaty. Modulated logarithmic Sobolev inequalities and generation of chaos. *Ann. Fac. Sci. Toulouse, Math. (6)*, 2023. arXiv: 2307.07587 [math.PR]. Forthcoming.

[201] Matthew Rosenzweig and Sylvia Serfaty. Relative entropy and modulated free energy without confinement via self-similar transformation. 2024. arXiv: 2402.13977 [math.AP].

[202] Grant Rotskoff, Samy Jelassi, Joan Bruna and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5508–5517. PMLR, 2019. URL: https://proceedings.mlr.press/v97/rotskoff19a.html.

[203] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Commun. Pure Appl. Math.*, 75(9):1889–1935, 2022. ISSN: 0010-3640. DOI: 10.1002/cpa.22074.

[204] Sebastian Ruder. An overview of gradient descent optimization algorithms. 2017. arXiv: 1609.04747 [cs.LG].

[205] Laure Saint-Raymond. *Des points vortex aux équations de Navier-Stokes [d'après P.-E. Jabin et Z. Wang]*. French. In *Sémin. Bourbaki, volume 2017/2018, exposés 1136-1150*. Volume 414. Astérique. 2019, Exposé nº 1143, pages 289–308. ISBN: 978-2-85629-915-9. DOI: 10.24033/ast.1087.

[206] Katharina Schuh. Global contractivity for Langevin dynamics with distribution-dependent forces and uniform in time propagation of chaos. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 60(2):753–789, 2024. ISSN: 0246-0203. DOI: 10.1214/22-AIHP1337.

[207] Laurent Schwartz. *Théorie des distributions. Nouvelle édition, entièrement corrigée, refondue et augmentée*. French, volume 9-10 of *Publ. Inst. Math. Univ. Strasbourg*. Hermann, Paris, 1966.

[208] Othmane Sebbouh, Robert M. Gower and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3935–3971. PMLR, 2021. URL: https://proceedings.mlr.press/v134/sebbouh21a.html.

[209] Sylvia Serfaty. Mean field limit for Coulomb-type flows. *Duke Math. J.*, 169(15):2887–2935, 2020. ISSN: 0012-7094. DOI: 10.1215/00127094-2020-0019. Appendix by Mitia Duerinckx and Sylvia Serfaty.

[210] Lloyd S. Shapley. *Some topics in two-person games*. In *Advances in Game Theory*. Volume 52. Ann. Math. Stud. 1964, pages 1–28.

[211] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Math. Oper. Res.*, 47(1):120–152, 2022. ISSN: 0364-765X. DOI: 10.1287/moor.2020.1118.

[212] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM J. Appl. Math.*, 80(2):725–752, 2020. ISSN: 0036-1399. DOI: 10.1137/18M1192184.

[213] Ilya Sutskever, James Martens, George Dahl and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA. PMLR, 2013. URL: https://proceedings.mlr.press/v28/sutskever13.html.

[214] Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *7th International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=H1ebTsActm.

[215] Taiji Suzuki, Atsushi Nitanda and Denny Wu. Uniform-in-time propagation of chaos for the mean-field gradient Langevin dynamics. In *11th International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=_JScUk9TBUn.

[216] Alain-Sol Sznitman. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*. Volume 1464, Lect. Notes Math. Pages 165–251. Springer, Berlin, 1991. ISBN: 3-540-53841-0. DOI: 10.1007/BFb0085169.

[217] Denis Talay. Stochastic Hamiltonian systems: Exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. *Markov Process. Relat. Fields*, 8(2):163–198, 2002. ISSN: 1024-2953.

[218] Alvin Tsz Ho Tse. *Quantitative propagation of chaos of McKean–Vlasov equations via the master equation*. PhD thesis, The University of Edinburgh, 2019. URL: http://hdl.handle.net/1842/36096.

[219] Belinda Tzen and Maxim Raginsky. A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled McKean–Vlasov dynamics. 2020. arXiv: 2002.01987 [cs.LG].

[220] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Adv. Neural Inf. Process. Syst.* Volume 32, pages 8094–8106. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf.

[221] Cédric Villani. *Hypocoercivity*, volume 950 of *Mem. Am. Math. Soc.* American Mathematical Society, Providence, RI, 2009. ISBN: 978-0-8218-4498-4; 978-1-4704-0564-9. DOI: 10.1090/S0065-9266-09-00567-5.

[222] Cédric Villani. *Optimal transport. Old and new*, volume 338 of *Grundlehren Math. Wiss.* Springer-Verlag, Berlin etc., 2009. ISBN: 978-3-540-71049-3. DOI: 10.1007/978-3-540-71050-9.

[223] Feng-Yu Wang. Criteria of spectral gap for Markov operators. *J. Funct. Anal.*, 266(4):2137–2152, 2014. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2013.11.016.

[224] Feng-Yu Wang. Distribution dependent SDEs for Landau type equations. *Stochastic Process. Appl.*, 128(2):595–621, 2018. ISSN: 0304-4149. DOI: 10.1016/j.spa.2017.05.006.

[225] Feng-Yu Wang. Exponential contraction in Wasserstein distances for diffusion semigroups with negative curvature. *Potential Anal.*, 53(3):1123–1144, 2020. ISSN: 0926-2601. DOI: 10.1007/s11118-019-09800-z.

[226] Feng-Yu Wang. Functional inequalities for non-symmetric stochastic differential equations. Chinese. *Sci. Sin., Math.*, 50(1):167–178, 2020. ISSN: 1674-7216. DOI: 10.1360/N012018-00150.

[227] Feng-Yu Wang. Hypercontractivity and applications for stochastic Hamiltonian systems. *J. Funct. Anal.*, 272(12):5360–5383, 2017. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2017.03.015.

[228] Feng-Yu Wang. Log-Sobolev inequalities: Different roles of Ric and Hess. *Ann. Probab.*, 37(4):1587–1604, 2009. ISSN: 0091-1798. DOI: 10.1214/08-AOP444.

[229] Songbo Wang. Sharp local propagation of chaos for mean field particles with $W^{-1,\infty}$ kernels. 2024. arXiv: 2403.13161 [math.PR].

[230] Songbo Wang. Uniform log-Sobolev inequalities for mean field particles with flat-convex energy. 2024. arXiv: 2408.03283 [math.PR].

[231] Colin Wei, Jason D. Lee, Qiang Liu and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Adv. Neural Inf. Process. Syst.* Volume 32, pages 9717–9724. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/8744cf92c88433f8cb04a02e6db69a0d-Paper.pdf.

[232] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman and Zhanxing Zhu. On the noisy gradient descent that generalizes as SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10367–10376. PMLR, 2020.

[233] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11(88):2543–2596, 2010. URL: http://jmlr.org/papers/v11/xiao10a.html.

[234] Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou and Tuo Zhao. Toward understanding the importance of noise in training neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7594–7602. PMLR, 2019.

[235] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7654–7663. PMLR, 2019. URL: https://proceedings.mlr.press/v97/zhu19e.html.

Remerciements

Je tiens à exprimer ma profonde gratitude à mes directeurs de thèse, Zhenjie Ren et Nizar Touzi. Leurs conseils avisés, leur soutien constant et la liberté qu'ils m'ont accordée ont été déterminants tout au long de mes recherches.

Je remercie sincèrement les rapporteurs de cette thèse, François Delarue et Daniel Lacker, pour le temps qu'ils ont consacré à une lecture attentive et rigoureuse de mon travail.

Je suis honoré que Pierre Cardaliaguet, Arnaud Guillin, Sylvie Méléard et Huyên Pham aient accepté de faire partie du jury, et je les remercie pour leur intérêt ainsi que pour leur participation à la soutenance.

Je remercie vivement Thierry Bodineau pour sa présence à la soutenance, pour l'encadrement de mon stage à l'époque du master, et pour les échanges plus récents qui ont révélé une convergence inattendue de nos intérêts de recherche.

Je remercie également les responsables de l'école doctorale, Olivier Le Maître et Stéphane Nonnenmacher, pour leur soutien et leur communication constante lors des difficultés administratives rencontrées durant mon doctorat.

J'ai beaucoup apprécié la qualité et la diversité de mes échanges avec mes collaborateurs, qui ont souvent suscité de nouvelles idées et ouvert de nouvelles perspectives. J'espère sincèrement que cette belle collaboration saura se prolonger et continuer à porter ses fruits à l'avenir, au bénéfice de nos réflexions communes.

Je n'oublie pas les amis et collègues rencontrés durant cette période, que ce soit à l'X, à Dauphine, à Jussieu, à l'IHP, à l'IHÉS ou ailleurs, avec qui j'ai partagé cette aventure et tant de moments précieux.

Enfin, et surtout, je suis profondément reconnaissant envers ma famille, mes amis proches et ma copine pour leur affection, leur patience, leur présence et leur compréhension à toute épreuve.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Comportements en temps long des dynamiques avec des interactions de champ moyen

Mots-clés : comportement en temps long, interaction de champ moyen, propagation du chaos, flot de gradient, entropie, inégalité de Sobolev logarithmique

Résumé : Cette thèse est consacrée à l'étude des comportements en temps long des dynamiques avec des interactions de champ moyen et des systèmes de particules associés. Pour la plupart des cas traités dans la thèse, la condition structurelle pour les comportements en temps long est la convexité plate de la fonctionnelle d'énergie de champ moyen, qui est différente de la convexité de déplacement étudiée dans les travaux classiques de transport optimal et de flot de gradient. La thèse est composée de trois parties. Dans la première partie, nous étudions les dynamiques de Langevin de champ moyen suramortie et sousamortie, qui sont des dynamiques de gradient associées à une fonctionnelle d'énergie libre de champ moyen, et nous montrons qu'elles présentent des propriétés de propagation du chaos uniforme en temps en exploitant leurs structures de gradient et une inégalité de Sobolev logarithmique uniforme. Dans la deuxième partie, nous développons d'abord quelques résultats techniques sur les inégalités de Sobolev logarithmiques et nous les appliquons pour obtenir la propagation du chaos uniforme en temps pour de diverses diffusions de McKean-Vlasov. En particulier, pour le modèle de vortex visqueux en 2D, nous développons des bornes de régularité fortes sur sa limite de champ moyen sur l'espace entier et nous montrons sa propagation du chaos par la méthode de Jabin-Wang ; nous étudions également son problème de taille du chaos en utilisant l'approche entropique de Lacker et nous obtenons des bornes optimales et uniformes en temps dans le régime de haute viscosité. Dans la dernière partie de la thèse, nous explorons d'autres dynamiques de champ moyen qui proviennent de problèmes d'optimisation convexes. Pour l'optimisation régularisée par l'entropie, nous étudions une dynamique d'auto-jeu fictif et une diffusion auto-interagissante et nous montrons leurs convergences en temps long vers la solution du problème d'optimisation. Nous considérons également un semigroupe de Schrödinger non linéaire, qui est un flot de gradient pour le problème d'optimisation régularisé par l'information de Fisher, et nous montrons sa convergence exponentielle sous une condition de trou spectral uniforme.

Title: Long-time behaviors of dynamics with mean field interactions

Keywords: long-time behavior, mean field interaction, propagation of chaos, gradient flow, entropy, logarithmic Sobolev inequality

Abstract: This thesis is devoted to the study of the long-time behaviors of dynamics with mean field interactions and their associated particle systems. For most cases treated in the thesis, the structural condition for the long-time behaviors is the flat convexity of the mean field energy functional, which is different from the displacement convexity studied in the classical works of optimal transport and gradient flow. The thesis is comprised of three parts. In the first part, we study the overdamped and underdamped mean field Langevin dynamics, which are gradient dynamics associated to a mean field free energy functional, and show their time-uniform propagation of chaos properties by exploiting their gradient structures and a uniform logarithmic Sobolev inequality. In the second part, we first develop some technical results on logarithmic Sobolev inequalities and apply them to get the time-uniform propagation of chaos for various McKean–Vlasov diffusions. Specifically, for the 2D viscous vortex model, we develop strong regularity bounds on its mean field limit on the whole space and show its propagation of chaos by the Jabin–Wang method; we also study its size of chaos problem using the entropy approach of Lacker and obtain time-uniform sharp bounds in the high viscosity regime. In the last part of the thesis, we explore alternative mean field dynamics that originate from convex optimization problems. For the entropy-regularized optimization, we study a fictitious self-play dynamics and a self-interacting diffusion and show their long-time convergences to the solution of the optimization problem. We also consider a non-linear Schrödinger semigroup, which is a gradient flow for the optimization problem regularized by Fisher information, and show its exponential convergence under a uniform spectral gap condition.

Institut polytechnique de Paris
91120 Palaiseau, France